Estimation and Inference for Set-identified Parameters Using Posterior Lower Probability

Toru Kitagawa*

Department of Economics, University College London

28, July, 2012

Abstract

In inference for set-identified parameters, Bayesian probability statements about unknown parameters do not coincide, even asymptotically, with frequentist's confidence statements. This paper aims to smooth out this disagreement from a robust Bayes perspective. I show that a class of prior distributions exists, with which the posterior inference statements drawn via the lower envelope (lower probability) of the class of posterior distributions asymptotically agrees with frequentist confidence statements for the identified set. With this class of priors, the statistical decision problems, including the point and set estimation of the set-identified parameters, are analyzed under the posterior gamma-minimax criterion.

Keywords: Partial Identification, Bayesian Robustness, Belief Function, Imprecise Probability, Gamma-minimax, Random Set.

*Email: t.kitagawa@ucl.ac.uk. I thank Gary Chamberlain, Andrew Chesher, Siddhartha Chib, Larry Epstein, Jean-Pierre Florens, Guido Imbens, Hiro Kaido, Charles Manski, Ulrich Müller, Andriy Norets, Adam Rosen, Kevin Song, and Elie Tamer for their valuable discussions and comments. I also thank the seminar participants at Academia Sinica, Brown, Cornell, Cowles Conference 2011, EC^2 2010, Harvard/MIT, Midwest Econometrics 2010, Northwestern, NYU, RES Conference 2011, Seoul National University, Simon Fraser University, and UBC for their helpful comments. All remaining errors are mine. Financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) is gratefully acknowledged.

1 Introduction

In inferring *identified* parameters in a parametric setup, the Bayesian probability statements about unknown parameters are found to be similar, at least asymptotically, to the frequentist confidence statements about the true value of the parameters. In partial identification analyses initiated by Manski (1989, 1990, 2003, 2007), such asymptotic harmony between the two inference paradigms breaks down (Moon and Schorfheide (2011)). The Bayesian interval estimates for the set-identified parameter are shorter, even asymptotically, than the frequentist ones, and they asymptotically lie inside the frequentist confidence intervals. Frequentists might interpret this phenomenon, Bayesian over-confidence in their inferential statements, as being fictitious. Bayesians, on the other hand, might consider that the frequentist confidence statements, which apparently lack posterior probability interpretation, raise some interpretative difficulty once data are observed.

The primary aim of this paper is to smooth out the disagreement between the two schools of statistical inference by applying the perspective of a robust Bayes inference, where one can incorporate partial prior knowledge into posterior inference. While there is a variety of robust Bayes approaches, this paper focuses on a *multiple prior* Bayes analysis, where the partial prior knowledge, or the robustness concern against prior misspecification, is modeled with a class of priors (ambiguous belief). The Bayes rule is applied to each prior to form a class of posteriors. The posterior inference procedures considered in this paper operate on the class of posteriors by focusing on their lower and upper envelopes, the so-called *posterior lower and upper probabilities*.

When the parameters are not identified, the prior distribution of the model parameters can be decomposed into two components: one that can be updated by data (*revisable prior knowledge*) and one that can never be updated by data (*unrevisable prior knowledge*). Given that the ultimate goal of the partially identification analysis is to establish a "domain of consensus" (Manski (2007)) among the set of assumptions that data are silent about, a natural way to incorporate this agenda into the robust Bayes framework is to design a prior class in such a way that it shares a single prior distribution for the revisable prior knowledge, but allows for arbitrary prior distributions for the unrevisable prior knowledge. Using this prior class as a prior input, this paper derives the posterior lower probability and investigates its analytical property. For an interval-identified parameter case, I also examine whether the inferential statements drawn via the posterior lower probability can asymptotically have any type of valid frequentist coverage probability in the partially identified setting.

Another question this paper examines is, with such class of priors, how to formulate and solve statistical decision problems including point estimation of the set-identified parameters. I approach this question by adapting the posterior gamma-minimax analysis, which can be seen as a minimax analysis with the multiple posteriors, and demonstrate that the proposed prior class leads to an analytically tractable and numerically solvable formulation of the posterior gamma-minimax decision problem, provided that the identified set for the parameter of interest can be computed for each possible distribution of data.

1.1 Related Literature

Estimation and inference in partially identified models are a growing research area in the field of econometrics. From the frequentist perspective, Horowitz and Manski (2000) construct confidence intervals for an interval identified set. Imbens and Manski (2004) propose uniformly asymptotically valid confidence sets for an interval-identified parameter, which are further extended by Stoye (2009). Chernozhukov, Hong, and Tamer (2007) develop a way to construct asymptotically valid confidence sets for an identified set based on the criterion function approach, which can be applied to a wide range of partially identified models including moment inequality models. In relation to the criterion function approach, the literature on the construction of confidence sets by inverting test statistics includes, but is not limited to, Andrews and Guggenberger (2009), Andrews and Soares (2010), and Romano and Shaikh (2010).

From the Bayesian perspective, Neath and Samaniego (1997), Poirier (1998), and Gustafson (2009, 2010) analyze how Bayesian updating performs when a model lacks identification. Liao and Jiang (2010) conduct a Bayesian inference for moment inequality models, based on the pseudo-likelihood. Moon and Schorfheide (2011) compare the asymptotic properties of frequentist and Bayesian inferences for set-identified models. My robust Bayes analysis is motivated by Moon and Schorfheide's important findings on the asymptotic disagreement between the frequentist and Bayesian inferences. Epstein and Seo (2012) focus on a set-identified model of entry games with multiple equilibria, and provide an axiomatic argument that justifies a single-prior Bayesian inference for a set-identified parameter. The current paper does not intend to provide any normative argument as to whether one should proceed with a single prior or multiple priors in inferring non-identified parameters.

The analysis of lower and upper probabilities originates with Dempster (1966, 1967a, 1967b, 1968), in his fiducial argument of drawing posterior inferences without specifying a prior distribution. The influence of Dempster's appears in the belief function analysis of Shafer (1976, 1982) and the imprecise probability analysis of Walley (1991). In the context of robust Bayes analysis, the lower and upper probabilities have been playing important roles in measuring the global sensitivity of the posterior (Berger (1984), Berger and Berliner (1986)) and also in characterizing a class of priors/posteriors (DeRobertis and Hartigan (1981), Wasserman (1989, 1990), and Wasserman and Kadane (1990)). In econometrics, pioneering work using multiple priors was carried out by Chamberlain and Leamer (1976), and Leamer (1982), who obtained the bounds for the posterior mean of the regression coefficients when a prior varies over a certain class. All of these previous studies did not explicitly consider non-identified models. This paper, in contrast, focuses on non-identified models, and aims to clarify a link between the early idea of the lower and upper probabilities and a recent issue on inferences in set-identified models.

The posterior lower probability to be obtained in this paper is an infinite-order monotone capacity, or equivalently, a containment functional in the random set theory. Beresteanu and Molinari (2008) and Beresteanu, Molchanov, and Molinari (2012) show the usefulness and wide applicability of the random set theory to a class of partially identified models by viewing observations as random sets, and the estimand (identified set) as its Aumann expectation. They propose an asymptotically valid frequentist inference procedure for the identified set by employing the central limit theorem applicable to the properly defined sum of random sets. Galichon and Henry (2006, 2009) and Beresteanu, Molchanov, and Molinari (2011) propose a use of infinite-order capacity in defining and inferring the identified set in the structural econometric model with multiple equilibria. The robust Bayes analysis of this paper closely relates to the literature of non-additive measures and random sets, but the way that these theories enter to the analysis differs from these previous works in the following ways. First, the class of models to be considered is assumed to have well-defined likelihood functions, and the lack of identification is modeled in terms of the "data-independent flat regions" of the likelihood. Ambiguity is not explicitly modeled at the level of observations, but instead ambiguity for the parameters is introduced through the absence of prior knowledge on each flat region of the likelihood. Second, I obtain the identified set as random sets, whose probability law is represented by the posterior lower probability. Here, the source of probability that induces the random identified set is the posterior uncertainty for the identifiable parameters, not the sampling probability of the observations. Third, the inferential statements to be proposed in the paper are made conditional on data, and they do not invoke any large-sample approximations.

The decision theoretic analysis in this paper employs the posterior gamma-minimax criterion, which leads to a decision that minimizes the worst case posterior risk over the class of posteriors. The gamma-minimax decision analysis often becomes challenging, both analytically and numerically, and the existing analyses are limited to rather simple parametric models with a certain choice of prior class (Betro and Ruggeri (1992), Chamberlain (2000), and Vidakovic (2000)). The specified prior class, in contrast, offers a general and feasible way to solve the posterior gamma-minimax decision problem, provided that the identified set for the parameter of interest can be computed for each of the identified parameter values. In a recent study by Song (2012), point estimation for an interval-identified parameter from the local asymptotic minimax approach is considered.

1.2 Plan of the Paper

The rest of the paper is organized as follows. In Section 2, the main results of this paper are presented using a simple example of missing data. Section 3 introduces the general framework, where I construct a class of prior distributions that can contain arbitrary unrevisable prior knowledge. I then derive the posterior lower and upper probabilities. Statistical decision analyses with multiple priors are examined in Section 4. In Section 5, how to construct the posterior credible regions based on the posterior lower probability is discussed and their large-sample behaviors are examined in an interval-identified parameter case. Proofs and lemmas are provided in Appendix A.

2 An Illustration: A Missing Data Example

This section illustrates the main results of the paper using an example of missing data (Manski (1989)). Let $Y \in \{1, 0\}$ be the binary outcome of interest (e.g., a worker is employed or not). Let $W \in \{1, 0\}$ be an indicator of whether Y is observed ($W = 1$) or not ($W = 0$), (i.e., the subject responded or not). Data are given by a random sample of size n, $x = \{(Y_i W_i, W_i) : i = 1, \ldots, n\}$.

The starting point of the analysis is to specify a parameter vector $\theta \in \Theta$ that pins down the distribution of the data and the parameter of interest. Here, θ can be specified by a vector of four probability masses: (θ_{yd}), $\theta_{yd} \equiv \Pr(Y = y, D = d)$, $y = 1, 0$, and $d = 1, 0$. The observed data likelihood for θ is written as

$$p(x|\theta) = \theta_{11}^{n_{11}} \theta_{01}^{n_{01}} \left[\theta_{10} + \theta_{00}\right]^{n_{mis}},$$

where $n_{11} = \sum_{i=1}^{n} Y_i W_i$, $n_{01} = \sum_{i=1}^{n} (1 - Y_i) W_i$, $n_{mis} = \sum_{i=1}^{n} (1 - W_i)$. This likelihood function depends on θ only through the three probability masses, $\phi = (\phi_{11}, \phi_{01}, \phi_{mis}) \equiv (\theta_{11}, \theta_{01}, \theta_{10} + \theta_{00})$, $\phi \in \Phi$, no matter what the observations are, so that the likelihood for θ has "data-independent" flat regions, which are expressed as a set-valued map of ϕ,

$$\Gamma(\phi) \equiv \{\theta \in \Theta : \theta_{11} = \phi_{11},\ \theta_{01} = \phi_{01},\ \theta_{10} + \theta_{00} = \phi_{mis}\}.$$

The parameter of interest is the mean of Y, which is written as a function of θ, $\eta \equiv \Pr(Y = 1) = \theta_{11} + \theta_{10}$. The identified set of η, $H(\phi)$, as the set-valued map of ϕ is defined by the range of η when θ varies over $\Gamma(\phi)$,

$$H(\phi) = [\phi_{11}, \phi_{11} + \phi_{mis}],$$

which are the Manski (1989)'s bounds of $\Pr(Y = 1)$.

The standard Bayes inference for η would proceed as follows; specify a prior of θ, update it using the Bayes rule, and marginalize the posterior of θ to η. If the likelihood for θ has data-independent flat regions as represented by $\{\Gamma(\phi) : \phi \in \Phi\}$, then the prior for θ conditional on $\{\theta \in \Gamma(\phi)\}$ (i.e., belief on how the proportion of missing observations, ϕ_{mis}, is divided into θ_{10} and θ_{00}) will never be updated by data. Consequently, the posterior of θ and possibly that of η become sensitive to the specification of such conditional priors of θ given ϕ. The robust Bayes procedure considered in this paper aims to make the posterior inference free from such sensitivity concerns by introducing multiple priors for θ. The way to construct a prior class is as follows. I first specify a single prior μ_ϕ for the identified parameters ϕ. In view of θ, prior μ_ϕ specifies how much prior belief should be assigned to each flat region of the θ's likelihood $\Gamma(\phi)$, whereas, depending on ways to allocate the assigned belief over $\theta \in \Gamma(\phi)$ (for each ϕ), the implied prior for θ may differ. Therefore, by collecting all the possible ways of allocate the assigned belief over $\{\theta \in \Gamma(\phi)\}$ for each ϕ, I can construct the following class of prior distributions of θ, $\mathcal{M}(\mu_\phi) = \left\{\mu_\theta : \mu_\theta(\Gamma(B)) = \mu_\phi(B) \quad \text{for all } B \subset \Phi\right\}$, where μ_θ denotes a prior distribution for θ.

By applying the Bayes rule to each $\mu_\theta \in \mathcal{M}(\mu_\phi)$ and marginalizing each posterior of θ for η, I obtain the class of posteriors of η, $\left\{F_{\eta|X} : \mu_\theta \in \mathcal{M}(\mu_\phi)\right\}$. I now summarize the class of posteriors of η by its lower envelope (lower probability), $F_{\eta|X*}(D) = \inf_{\mu_\theta \in \mathcal{M}(\mu_\phi)} F_{\eta|X}(D)$, which maps subset D in the parameter space of η to $[0,1]$. In words, the posterior lower probability evaluated at D says that the posterior belief allocated for $\{\eta \in D\}$ is at least $F_{\eta|X*}(D)$, no matter which $\mu_\theta \in \mathcal{M}(\mu_\phi)$ is used.

The main theorem of this paper shows that the posterior lower probability satisfies

$$F_{\eta|X*}(D) = F_{\phi|X}(\{\phi : H(\phi) \subset D\}),$$

where $F_{\phi|X}$ denotes the posterior distribution of ϕ implied from the prior μ_ϕ. The key insight of this equality is that, with prior class $\mathcal{M}(\mu_\phi)$, drawing inference for η based on its posterior lower probability is done by analyzing the probability law of random sets $H(\phi)$, $\phi \sim F_{\phi|X}$.

Leaving their formal analysis to the later sections of this paper, I now outline the implementation of the posterior lower probability inference for η proposed in this paper.

1. Specify a prior for ϕ and update it by the Bayes rule. When a credible prior for ϕ is not available, a reasonably "non-informative" prior may be used as far as the posterior of ϕ is proper.[1]

2. Let $\{\phi_s : s = 1, \ldots, S\}$ be random draws of ϕ from the posterior of ϕ. The mean and median of the posterior lower probability of η can be defined via the gamma-minimax

[1] See Kass and Wasserman (1996) for a survey of "reasonably" non-informative priors.

decision criterion, and they can be approximated by

$$\arg\min_{a} \frac{1}{S}\sum_{s=1}^{S} \sup_{\eta\in H(\phi_s)} (a-\eta)^2 \quad \text{and} \quad \arg\min_{a} \frac{1}{S}\sum_{s=1}^{S} \sup_{\eta\in H(\phi_s)} |a-\eta|,$$

respectively.

3. The posterior lower credible region of η at credibility level $1-\alpha$, which can be interpreted as a $(1-\alpha)$- level set of the posterior lower probability of η, is defined by the smallest interval that contains $H(\phi)$ with posterior probability $1-\alpha$ (Proposition 5.1 in this paper proposes an algorithm to compute the posterior lower credible region for interval-identified cases). Under certain regularity conditions that are satisfied in the current missing data example, the posterior lower credible region of η asymptotically attains the frequentist coverage probability $1-\alpha$ for the true identified *set* $H(\phi_0)$. where ϕ_0 is the value of ϕ corresponding to the sampling distribution of data.

3 Multiple-prior Analysis and the Lower and Upper Probabilities

3.1 Likelihood and Set Identification: The General Framework

Let $(\mathbf{X},\mathcal{X})$ and $(\Theta,\mathcal{A})$ be measurable spaces of a sample $X\in\mathbf{X}$ and a parameter vector $\theta\in\Theta$, respectively. The analytical framework of this paper covers both a parametric model $\Theta=\mathcal{R}^d$, $d<\infty$, and a non-parametric model where Θ is a separable Banach space. The sample size is implicit in the notation. Let μ_θ be a marginal probability distribution on the parameter space $(\Theta,\mathcal{A})$, referred to as a *prior distribution* for θ. Assume that the conditional distribution of X given θ exists and has the probability density $p(x|\theta)$ at every $\theta\in\Theta$ with respect to a σ-finite measure on $(\mathbf{X},\mathcal{X})$.

The parameter vector θ may consist of parameters that determine the behaviors of the economic agents, as well as those that characterize the distribution of the unobserved heterogeneities in the population. In the context of the missing data or counterfactual causal models, θ indexes the distribution of the underlying population outcomes or the potential outcomes. In all of these cases, the parameter θ should be distinguished from the parameters that are solely used to index the sampling distribution of observations. The identification problem of θ typically arises in this context. If multiple values of θ generate the same distribution of data, then these θ's are *observationally equivalent* and the identification of θ fails. In terms of the likelihood function $p(x|\theta)$, the observational equivalence of θ and $\theta' \neq \theta$ means that the values of the likelihood at θ and θ' are equal for every possible sample, i.e., $p(x|\theta) = p(x|\theta')$ for every $x \in \mathbf{X}$ (Rothenberg (1971), Drèze (1974), and Kadane (1974)). I represent the observational equivalence relation of θ's by a many-to-one function $g : (\Theta, \mathcal{A}) \to (\Phi, \mathcal{B})$:

$$g(\theta) = g(\theta') \text{ if and only if } p(x|\theta) = p(x|\theta') \text{ for all } x \in \mathbf{X}.$$

The equivalence relationship partitions the parameter space Θ into equivalent classes, in each of which the likelihood of θ is "flat", irrespective of observations, and $\phi = g(\theta)$ maps each of these equivalent classes to a point in another parameter space Φ. In the language of structural models in econometrics (Hurwicz (1950), and Koopman and Reiersol (1950)), $\phi = g(\theta)$ is interpreted as the reduced-form parameter that carries all the information for the structural parameters θ through the value of the likelihood function. In the literature of Bayesian statistics, $\phi = g(\theta)$ is referred to as the *minimally sufficient parameter* (*sufficient parameter* for short), and the range space of $g(\cdot)$, $(\Phi, \mathcal{B})$, is called the *sufficient parameter space* (Barankin (1960), Dawid (1979), Florens and Mouchart (1977), Picci (1977), and Florens, Mouchart, and Rolin (1990)).[2]

In the presence of sufficient parameters, the likelihood depends on θ, only through the function $g(\theta)$, i.e., there exists a $\mathcal{B}$-measurable function $\hat{p}(x|\cdot)$ such that

$$p(x|\theta) = \hat{p}(x|g(\theta)) \quad \forall x \in \mathbf{X} \text{ and } \theta \in \Theta \tag{3.1}$$

holds (Lemma 2.3.1 of Lehmann and Romano (2005)).

Denote the inverse image of $g(\cdot)$ by Γ:

$$\Gamma(\phi) = \{\theta \in \Theta : g(\theta) = \phi\},$$

[2] Florens and Simoni (2011) provide comprehensive discussions on the relationship between frequentist and Bayesian identification.

where $\Gamma(\phi)$ and $\Gamma(\phi')$ for $\phi \neq \phi'$ are disjoint, and $\{\Gamma(\phi) \ ; \ \phi \in \Phi\}$ constitutes a partition of Θ. I assume $g(\Theta) = \Phi$, so $\Gamma(\phi)$ is non-empty for every $\phi \in \Phi$.[3]

In the set-identified model, the parameter of interest $\eta \in \mathcal{H}$ is a subvector or a transformation of θ denoted by $\eta = h(\theta)$, $h : (\Theta, \mathcal{A}) \to (\mathcal{H}, \mathcal{D})$. The formal definition of the identified set of η is given as follows.

Definition 3.1 (Identified Set of η)
(i) The identified set of η is a set-valued map $H : \Phi \rightrightarrows H$ defined by the projection of $\Gamma(\phi)$ onto H through $h(\cdot)$, $H(\phi) \equiv \{h(\theta) : \theta \in \Gamma(\phi)\}$.

(ii) The parameter $\eta = h(\theta)$ is point-identified at ϕ if $H(\phi)$ is a singleton, and η is set-identified at ϕ if $H(\phi)$ is not a singleton.

Note that the identification of η is defined in the pre-posterior sense because it is based on the likelihood evaluated at every possible realization of a sample, not only for the observed one.

3.2 Examples

I now provide some examples, in addition to the illustrating example of Section 2, both to illustrate the above concepts and notations, and to provide a concrete focus for the later development.

Example 3.1 (Bounding ATE by Linear Programming) *Consider the treatment effect model with incompliance and a binary instrument $Z \in \{1, 0\}$, as considered in Imbens and Angrist (1994), and Angrist, Imbens, and Rubin (1996). Assume that the treatment status and the outcome of interest are both binary. Let $(W_1, W_0) \in \{1, 0\}^2$ be the potential treatment status in response to the instrument, and $W = ZW_1 + (1 - Z)W_0$ be the observed treatment status. $(Y_1, Y_0) \in \{1, 0\}^2$ is a pair of treated and control outcomes and*

[3] In an observationally restrictive model, in the sense of Koopman and Reiersol (1950), $\hat{p}(x|\cdot)$ likelihood function for the sufficient parameters, is well defined for a domain larger than $g(\Theta)$ (see Example 3.1 in Section 3.2). In this case, the model possesses the falisifiability property, and $\Gamma(\phi)$ can be empty for some $\phi \in \Phi$.

$Y = WY_1 + (1 - W)Y_0$ is the observed outcome. Data is a random sample of (Y_i, W_i, Z_i). Following Imbens and Angrist (1994), consider partitioning the population into four subpopulations defined in terms of the potential treatment-selection responses:

$$T_i = \begin{cases} c & \textit{if } \; W_{1i} = 1 \; \textit{ and } \; W_{0i} = 0 & : \textit{complier}, \\ at & \textit{if } \; W_{1i} = W_{0i} = 1 & : \textit{always-taker}, \\ nt & \textit{if } \; W_{1i} = W_{0i} = 0 & : \textit{never-taker}, \\ d & \textit{if } \; W_{1i} = 0 \; \textit{ and } \; W_{0i} = 1 & : \textit{defier}, \end{cases}$$

where T_i is the indicator for the types of selection responses.

Assume a randomized instrument, $Z \perp (Y_1, Y_0, W_1, W_0)$. Then, the distribution of observables and the distribution of potential outcomes satisfy the following equalities for $y \in \{1, 0\}$:

$$\begin{aligned} \Pr(Y = y, W = 1 | Z = 1) &= \Pr(Y_1 = y, T = c) + \Pr(Y_1 = y, T = at), \qquad (3.2) \\ \Pr(Y = y, W = 1 | Z = 0) &= \Pr(Y_1 = y, T = d) + \Pr(Y_1 = y, T = at), \\ \Pr(Y = y, W = 0 | Z = 1) &= \Pr(Y_0 = y, T = d) + \Pr(Y_1 = y, T = nt), \\ \Pr(Y = y, W = 0 | Z = 0) &= \Pr(Y_0 = y, T = c) + \Pr(Y_1 = y, T = nt). \end{aligned}$$

Ignoring the marginal distribution of Z, a full parameter vector of the model can be specified by a joint distribution of (Y_1, Y_0, T):

$$\theta = (\Pr(Y_1 = y, Y_0 = y', T = t) : y = 1, 0, \quad y' = 1, 0, \quad t = c, nt, at, d) \in \Theta,$$

where Θ is the 16-dimensional probability simplex. Let ATE be the parameter of interest.

$$\begin{aligned} \eta \;\; &\equiv \;\; E(Y_1 - Y_0) = \sum_{t=c,nt,at,d} [\Pr(Y_1 = 1, T = t) - \Pr(Y_0 = 1, T = t)] \\ &= \;\; \sum_{t=c,nt,at,d} \sum_{y=1,0} [\Pr(Y_1 = 1, Y_0 = y, T = t) - \Pr(Y_1 = y, Y_0 = 1, T = t)] \\ &\equiv \;\; h(\theta). \end{aligned}$$

The likelihood conditional on Z depends on θ only through the distribution of (Y, W) given Z, so the sufficient parameter vector consists of eight probability masses:

$$\phi = (\Pr(Y = y, W = w | Z = z) : y = 1, 0, \quad d = 1, 0, \quad z = 1, 0).$$

The set of equations (3.2) defines $\Gamma(\phi)$ the set of observationally equivalent distributions of (Y_1, Y_0, T), when the data distribution is given at ϕ. Balke and Pearl (1997) derive the identified set of ATE, $H(\phi) = h(\Gamma(\phi))$, by maximizing or minimizing $h(\theta)$, subject to $\theta \in \Theta$ and constraints (3.2). This optimization can be solved by linear programming and $H(\phi)$ is obtained as a convex interval.

Note that, in this model, special attention is needed for the sufficient parameter space Φ to ensure that $\Gamma(\phi)$ is non-empty. Pearl (1995) shows that the distribution of data is compatible with the instrument exogeneity condition, $Z \perp (Y_1, Y_0, W_1, W_0)$, if and only if

$$\max_w \sum_y \max_z \{\Pr(Y = y, W = w)|Z = z\} \leq 1. \tag{3.3}$$

This implies that in order to guarantee $\Gamma(\phi) \neq \emptyset$, a prior distribution for ϕ puts probability one for ϕ that fulfill (3.3).

Example 3.2 (Linear Moment Inequality Model) *Consider the model where the parameter of interest $\eta \in \mathcal{H}$ is characterized by linear moment inequalities,*

$$E(m(X) - A\eta) \geq 0,$$

where the parameter space $\mathcal{H}$ is a subset of $\mathcal{R}^L$, $m(X)$ is a J-dimensional vector of known functions of an observation, and A is a $J \times L$ known constant matrix. By augmenting the J-dimensional parameter $\lambda \in [0, \infty)^J$, these moment inequalities can be written as the J-moment equalities,[4]

$$E(m(X) - A\eta - \lambda) = 0.$$

To obtain a likelihood function for the current moment equality model, specify the full parameter vector to be $\theta = (\eta, \lambda) \in \mathcal{H} \times [0, \infty)^J$, and consider the exponentially tilted empirical likelihood for θ as considered in Schennach (2005). Let $x = (x_1, \ldots, x_n)$ be a size n random sample of observations, and define $g(\theta) = A\eta + \lambda$. If the convex hull of $\cup_i \{m(x_i) - g(\theta)\}$ contains the origin, then the exponentially tilted empirical likelihood is written as

$$p(x|\theta) = \prod w_i(\theta),$$

[4] I owe the Bayesian formulation of the moment inequality model shown here to Tony Lancaster (personal communication 2006).

where

$$w_i(\theta) = \frac{\exp\left\{\gamma(g(\theta))'\left(m(x_i) - g(\theta)\right)\right\}}{\sum_{i=1}^{n}\exp\left\{\gamma(g(\theta))'\left(m(x_i) - g(\theta)\right)\right\}},$$

$$\gamma(g(\theta)) = \arg\min_{\gamma \in R_+^J}\left\{\sum_{i=1}^{n}\exp\left\{\gamma'\left(m(x_i) - g(\theta)\right)\right\}\right\}.$$

Thus, the parameter $\theta = (\eta, \lambda)$ enters the likelihood only through $g(\theta) = A\eta + \lambda$. Consequently, I take $\phi = g(\theta)$ to be the sufficient parameters. The identified set for θ is given by

$$\Gamma(\phi) = \left\{(\eta, \lambda) \in \mathcal{H} \times [0, \infty)^L : A\eta + \lambda = \phi\right\}.$$

The coordinate projection of $\Gamma(\phi)$ onto $\mathcal{H}$ yields $H(\phi)$, the identified set for η (Bertsimas and Tsitsiklis (1997, Chap.2) for an algorithm for projecting a polyhedron).

3.3 Unrevisable Prior Knowledge and a Class of Priors

Let μ_θ be a prior of θ and μ_ϕ be the marginal probability measure on the sufficient parameter space $(\Phi, \mathcal{B})$ induced by μ_θ and $g(\cdot)$:

$$\mu_\phi(B) = \mu_\theta(\Gamma(B)) \quad \text{for all } B \in \mathcal{B}.$$

Let $x \in \mathbf{X}$ be sampled data. The posterior distribution of θ, denoted by $F_{\theta|X}(\cdot)$, is obtained as

$$F_{\theta|X}(A) = \int_\Phi \mu_{\theta|\phi}(A|\phi)dF_{\phi|X}(\phi), \qquad A \in \mathcal{A}, \tag{3.4}$$

where $\mu_{\theta|\phi}(A|\phi)$ denotes the conditional distribution of θ given ϕ, and $F_{\phi|X}(\cdot)$ is the posterior distribution of ϕ.

The posterior distribution of θ given in (3.4) shows that the prior distribution for θ marginalized to ϕ can be updated by data, while the conditional prior of θ given ϕ is never be updated by the data because the likelihood is flat on $\Gamma(\phi) \subset \Theta$ for any realizations of the sample. In this sense, the prior information marginalized to the sufficient parameter μ_ϕ can be interpreted as the *revisable prior knowledge*, and the conditional priors of θ given ϕ, $\left\{\mu_{\theta|\phi}(\cdot|\phi) : \phi \in \Phi\right\}$ can be interpreted as the *unrevisable prior knowledge*. If one wants to summarize the posterior uncertainty of θ in the form of a probability distribution on $(\Theta, \mathcal{A})$, as recommended in the Bayesian paradigm, he needs to have a single prior distribution of θ, which necessarily induces unique unrevisable prior knowledge $\mu_{\theta|\phi}$. If he could justify his choice of $\mu_{\theta|\phi}$ by any credible prior information, the standard Bayesian updating (3.4) would yield a valid posterior distribution of θ. A challenging situation would arise if one is short of a credible prior distribution of θ. In this case, the researcher, who is aware that $\mu_{\theta|\phi}$ will never be updated by data, might feel anxious in implementing the Bayesian inference procedure, because an unconfidently specified $\mu_{\theta|\phi}$ can have a significant influence to the subsequent posterior inference.

The robust Bayes analysis in this paper specifically focuses on such a situation, and introduce ambiguity for the conditional prior $\left\{\mu_{\theta|\phi}(\cdot|\phi) : \phi \in \Phi\right\}$ in the form of multiple priors. Specifically, given μ_ϕ a prior on $(\Phi, \mathcal{B})$ specified by the researcher, consider the class of prior distributions of θ defined by:

$$\mathcal{M}(\mu_\phi) = \left\{\mu_\theta : \mu_\theta(\Gamma(B)) = \mu_\phi(B) \text{ for every } B \in \mathcal{B}\right\}.$$

$\mathcal{M}(\mu_\phi)$ consists of prior distributions of θ whose marginal distribution for the sufficient parameters coincides with the prespecified μ_ϕ.[5] This paper proposes to use $\mathcal{M}(\mu_\phi)$ as a prior input for the posterior analysis, meaning that, with accepting to specify a single prior distribution for the sufficient parameters ϕ, I leave the conditional priors $\mu_{\theta|\phi}$ unspecified and allow for arbitrary ones as long as $\mu_\theta(\cdot) = \int_\Phi \mu_{\theta|\phi}(\cdot|\phi)d\mu_\phi$ yields a probability measure on $(\Theta, \mathcal{A})$.[6]

In the subsequent analysis, I shall not discuss how to select μ_ϕ, and shall treat μ_ϕ as given. The influence of μ_ϕ on the posterior of ϕ will diminish as the sample size increases, so the sensitivity issue of the posterior of ϕ is expected to be less severe when the sample size is moderate or large.

[5] I thank Jean-Pierre Florens for suggesting this representation of the prior class.

[6] Sufficient parameters ϕ are defined by examining the entire model $\{p(x|\theta) : x \in \mathcal{X}, \theta \in \Theta\}$, so that the prior class $\mathcal{M}(\mu_\phi)$ is, by construction, model dependent. This distinguishes the current approach from the standard robust Bayes analysis where a prior class represents the researcher's subjective assessment of his imprecise prior knowledge (Berger (1985)).

3.4 Posterior Lower and Upper Probabilities

The Bayes rule is applied to each prior in $\mathcal{M}(\mu_\phi)$ to generate the class of posterior distributions of θ. Consider summarizing the posterior class by the posterior lower probability $F_{\theta|X*}(\cdot) : \mathcal{A} \to [0,1]$ and the posterior upper probability $F^*_{\theta|X}(\cdot) : \mathcal{A} \to [0,1]$, defined as

$$\begin{aligned} F_{\theta|X*}(A) &\equiv \inf_{\mu_\theta \in \mathcal{M}(\mu_\phi)} F_{\theta|X}(A), \\ F^*_{\theta|X}(A) &\equiv \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} F_{\theta|X}(A). \end{aligned}$$

Note that the posterior lower probability and the upper probability have a conjugate property, $F_{\theta|X*}(A) = 1 - F^*_{\theta|X}(A^c)$, so it suffices to focus on one of them in deriving their analytical form. In order to obtain $F_{\theta|X*}(\cdot)$, the following regularity conditions are assumed.

Condition 3.1

(i) A prior for ϕ, μ_ϕ, is proper and absolutely continuous with respect to a σ-finite measure on $(\Phi, \mathcal{B})$.

(ii) $g : (\Theta, \mathcal{A}) \to (\Phi, \mathcal{B})$ is measurable and its inverse image $\Gamma(\phi)$ is a closed set in Θ, μ_ϕ-almost every $\phi \in \Phi$.

(iii) $h : (\Theta, \mathcal{A}) \to (\mathcal{H}, \mathcal{D})$ is measurable and $H(\phi) = h(\Gamma(\phi))$ is a closed set in H, μ_ϕ-almost every $\phi \in \Phi$.

These conditions are imposed for $\Gamma(\phi)$ and $H(\phi)$ to be interpreted as random closed sets induced by a probability measure on $(\Phi, \mathcal{B})$.[7] The closedness of $\Gamma(\phi)$ and $H(\phi)$ are implied, for instance, by continuity of $g(\cdot)$ and $h(\cdot)$.

Theorem 3.1 *Assume Condition 3.1.*

[7] The inference procedure proposed in this paper can be implemented as long as the posterior of ϕ is proper. However, how to accommodate an improper prior for ϕ in the development of the analytical results is beyond the scope of this paper.

(i) For each $A \in \mathcal{A}$,

$$F_{\theta|X*}(A) = F_{\phi|X}(\{\phi : \Gamma(\phi) \subset A\}), \tag{3.5}$$

$$F^*_{\theta|X}(A) = F_{\phi|X}\left(\{\phi : \Gamma(\phi) \cap A \neq \emptyset\}\right), \tag{3.6}$$

where $F_{\phi|X}(B)$, $B \in \mathcal{B}$, is the posterior probability measure of ϕ.

(ii) Define the posterior lower and upper probabilities of $\eta = h(\theta)$ by

$$F_{\eta|X*}(D) \equiv \inf_{\mu_\theta \in \mathcal{M}(\mu_\phi)} F_{\theta|X}(h^{-1}(D)),$$
$$F^*_{\eta|X}(D) \equiv \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} F_{\theta|X}(h^{-1}(D)), \text{ for } D \in \mathcal{D}.$$

It holds

$$F_{\eta|X*}(D) = F_{\phi|X}(\{\phi : H(\phi) \subset D\}),$$
$$F^*_{\eta|X}(D) = F_{\phi|X}(\{\phi : H(\phi) \cap D \neq \emptyset\}).$$

Proof. For a proof of (i), see Appendix A. For a proof of (ii), see equation (3.7). ■

The expression for $F_{\theta|X*}(A)$ implies that the posterior lower probability on A calculates the probability that the set $\Gamma(\phi)$ is contained in subset A in terms of the posterior probability law of ϕ. On the other hand, the upper probability is interpreted as the posterior probability that the set $\Gamma(\phi)$ hits subset A. The second statement of the theorem provides a procedure for marginalizing the lower and upper probabilities of θ into those of the parameter of interest η. The expressions of $F_{\eta|X*}(D)$ and $F^*_{\eta|X}(D)$ are simple and easy to interpret: the lower and upper probabilities of $\eta = h(\theta)$ are the containment and hitting probabilities of the random sets obtained by projecting $\Gamma(\phi)$ through $h(\cdot)$. This marginalization rule of the lower probability follows from

$$\begin{aligned} F_{\eta|X*}(D) &= F_{\theta|X*}(h^{-1}(D)) \\ &= F_{\phi|X}(\{\phi : \Gamma(\phi) \subset h^{-1}(D)\}) \\ &= F_{\phi|X}(\{\phi : H(\phi) \subset D\}). \end{aligned} \tag{3.7}$$

Note that, In the standard Bayesian inference, marginalization of the posterior of θ to η is conducted by integrating the posterior probability measure of θ for η, while in the lower probability inference, marginalization for η corresponds to projecting random sets $\Gamma(\phi)$ via $\eta = h(\cdot)$. This stark contrast between the standard Bayes and the multiple prior robust Bayes inference highlights how the introduction of ambiguity changes the way of eliminating the nuisance parameters in the posterior inference.

As is known in the literature (e.g., Huber (1973)), the lower probability of a set of probability measures is a monotone nonadditive measure (capacity). Furthermore, in the current specification of the prior class, the representation of the lower probability obtained in Theorem 3.1 implies that the resulting posterior lower and upper probabilities are *supermodular* and *submodular*, respectively.

Corollary 3.1 *Assume Condition 3.1. The posterior lower and upper probabilities of θ are supermodular and submodular, respectively. For A_1, $A_2 \in \mathcal{A}$ subsets in Θ,*

$$\begin{aligned} F_{\theta|X*}(A_1 \cup A_2) + F_{\theta|X*}(A_1 \cap A_2) &\geq F_{\theta|X*}(A_1) + F_{\theta|X*}(A_2), \\ F^*_{\theta|X}(A_1 \cup A_2) + F^*_{\theta|X}(A_1 \cap A_2) &\leq F^*_{\theta|X}(A_1) + F^*_{\theta|X}(A_2). \end{aligned}$$

Also, the posterior lower and upper probabilities of η are supermodular and submodular, respectively. For D_1, $D_2 \in \mathcal{D}$ subsets in $\mathcal{H}$,

$$\begin{aligned} F_{\eta|X*}(D_1 \cup D_2) + F_{\eta|X*}(D_1 \cap D_2) &\geq F_{\eta|X*}(D_1) + F_{\eta|X*}(D_2), \\ F^*_{\eta|X}(D_1 \cup D_2) + F^*_{\eta|X}(D_1 \cap D_2) &\leq F^*_{\eta|X}(D_1) + F^*_{\eta|X}(D_2). \end{aligned}$$

The results of Theorem 3.1 (i) can be seen as a special case of Wasserman's (1990) general construction of the posterior lower and upper probabilities. Whereas, one notable difference from Wasserman's analysis is that, with prior class $\mathcal{M}\left(\mu_\phi\right)$, the lower probability of the *posterior* class becomes an ∞-order monotone capacity (a containment functional of random sets).[8] This plays a crucial role in simplifying the gamma minimax analysis considered in the next section.

4 Posterior Gamma-minimax Analysis for $\eta = h(\theta)$

In the standard Bayesian posterior analysis, a statistical decision problem involving η (e.g., point estimation for η) is straightforward, minimizing the posterior risk. When the posterior information of η is summarized by the class of posteriors, how should the optimal statistical action be solved? This section studies this problem by adapting the posterior gamma-minimax analysis.

Let $a \in \mathcal{H}_a$ be an action, where $\mathcal{H}_a$ is an action space. In the case of the point estimation problem for η, an action is interpreted as reporting a non-randomized point estimate for η, where action space $\mathcal{H}_a$ is a subset of $\mathcal{H}$. Given an action a to be taken, and η being the true state of nature, a loss function $L(\eta, a) : \mathcal{H} \times \mathcal{H}_a \to \mathcal{R}_+$ yields the cost to the decision maker of taking action a. I assume that the loss function $L(\eta, a)$ is non-negative.

If a single prior for θ, μ_θ, were given, the *posterior risk* would be defined by

$$\rho(\mu_\theta, a) \equiv \int_{\mathcal{H}} L(\eta, a) dF_{\eta|X}(\eta), \tag{4.1}$$

where the first argument μ_θ in the posterior risk represents the dependence of the posterior of η on the specification of the prior for θ. Our analysis involves multiple priors, so the class of posterior risks $\left\{\rho(\mu_\theta, a) : \mu_\theta \in \mathcal{M}(\mu_\phi)\right\}$ is considered. The *posterior gamma-minimax criterion*[9] ranks actions in terms of the worst case posterior risk (*upper posterior risk*):

$$\rho^*(\mu_\phi, a) \equiv \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \rho(\mu_\theta, a),$$

[8] Wasserman (1990, p.463) posed an open question asking which class of priors can assure the posterior lower probability to be a containment functional of random sets. Theorem 3.1 provides an answer to his open question in the situation where the model lacks identifiability.

[9] In the robust Bayes literature, the class of prior distributions is often denoted by Γ. This is why it is called the gamma-minimax criterion. Unfortunately, in the literature of belief functions and lower and upper probabilities, Γ often denotes a set-valued mapping that generates the lower and upper probabilities. In this paper, we adopt the latter notational convention, but still refer to the decision criterion as the gamma-minimax criterion.

where the first argument in the upper posterior risk represents the dependence of the prior class on a prior for ϕ.

Definition 4.1 *A posterior gamma-minimax action a_x^* with respect to prior class $\mathcal{M}\left(\mu_\phi\right)$ is an action that minimizes the upper posterior risk, i.e.,*

$$\rho^*(\mu_\phi, a_x^*) = \inf_{a \in \mathcal{H}_a} \rho^*(\mu_\phi, a) = \inf_{a \in \mathcal{H}_a} \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \rho(\mu_\theta, a).$$

The gamma-minimax decision approach involves a favor for a conservative action that guards against the least favorable prior within the class, and it can be seen as a compromise of the Bayesian decision principle and the minimax decision principle.

The next proposition shows that the upper posterior risk $\rho^*(\mu_\phi, a)$ equals the Choquet expected loss with respect to the posterior upper probability.

Proposition 4.1 *Under Condition 3.1, the upper posterior risk satisfies*

$$\rho^*(\mu_\phi, a) = \int L(\eta, a) dF^*_{\eta|X}(\eta) = \int_\Phi \sup_{\eta \in H(\phi)} L(\eta, a) dF_{\phi|X}(\phi), \tag{4.2}$$

whenever $\int L(\eta, a) dF^*_{\eta|X}(\eta) < \infty$, where $\int L(\eta, a) dF^*_{\eta|X}(\eta)$ is the Choquet integral.

Proof. See Appendix A. ■

The third expression in (4.2) shows that the posterior gamma-minimax criterion is written as the expectation of the worst-case loss function, $\sup_{\eta \in H(\phi)} L(\eta, a)$, with respect to the posterior of ϕ. The supremum part stems from the ambiguity of η: given ϕ, what the researcher knows about η is only that it lies within the identified set $H(\phi)$, and, following the minimax principle, he forms the loss by supposing that the nature chooses the worst case in response to his/her action a. On the other hand, the expectation in ϕ represents the posterior uncertainty of the identified set $H(\phi)$: with the finite number of observations, the identified set of η is known with some uncertainty as summarized by the posterior of ϕ. The posterior gamma-minimax criterion combines such ambiguity of η with the posterior uncertainty of the identified set $H(\phi)$ to yield a single objective function to be minimized.[10]

[10] The posterior gamma minimax action a_x^* can be interpreted as a Bayes action for some *posterior* distributions in the class. For instance, in case of the quadratic loss, the saddle-point argument implies that the gamma-minimax action a_x^* corresponds to the mean of a posterior distribution (Bayes action) that has maximal posterior variance in the class.

Although a closed-form expression of a_x^* is not, in general, available, this proposition suggests a simple numerical algorithm for approximating a_x^* using a random sample of ϕ from its posterior $F_{\phi|X}$. Let $\{\phi_s\}_{s=1}^S$ be S random draws of ϕ from posterior $F_{\phi|X}$. Then, a_x^* can be approximated by

$$\hat{a}_x^* \equiv \arg\min_{a \in \mathcal{H}_a} \frac{1}{S} \sum_{s=1}^{S} \sup_{\eta \in H(\phi_s)} L(\eta, a).$$

The gamma minimax decisions are usually dynamically inconsistent; a *posteriori* optimal gamma-minimax action does not coincide with an unconditional optimal gamma-minimax decision. This is also the case with out prior class, and this will imply that a_x^* fails to be a Bayes decision with respect to any single prior in the class $\mathcal{M}\left(\mu_\phi\right)$. See Appendix B for an example and further discussion.

As an alternative to the posterior gamma-minimax action, the gamma-minimax *regret* criterion may be considered (Berger (1985, p. 218), and Rios Insua, Ruggeri, and Vidakovic (1995)). Appendix B provides some analytical results of the posterior gamma-minimax regret analysis where the parameter of interest η is a scalar and the loss function is quadratic, $L(\eta, a) = (\eta - a)^2$. There, it is shown that the posterior gamma-minimax regret decision can differ from the posterior gamma-minimax decision derived above, but that they converge to the same limit asymptotically.

5 Set Estimation of η

In the standard Bayesian inference, set estimation is often conducted by reporting the contour sets of the posterior probability density of η (highest posterior density region). If the posterior information for η is summarized by the lower and upper probabilities, how should we conduct set estimation of η?

5.1 Posterior Lower Credible Region

For $\alpha \in (0,1)$, consider a subset $C_{1-\alpha} \subset \mathcal{H}$ such that the posterior lower probability $F_{\eta|X*}(C_{1-\alpha})$ is greater than or equal to $1-\alpha$:

$$F_{\eta|X*}(C_{1-\alpha}) = F_{\phi|X}(H(\phi) \subset C_{1-\alpha})) \geq 1 - \alpha. \tag{5.1}$$

$C_{1-\alpha}$ is interpreted as "a set on which the posterior credibility of η is *at least* $1-\alpha$, *no matter which posterior is chosen within the class*". If I drop the italicized part from this statement, I obtain the usual interpretation of the posterior credible region, so $C_{1-\alpha}$ defined in this way seems to be a natural extension of the Bayesian posterior credible regions to those of the posterior lower probability. Analogous to the Bayesian posterior credible region, there are multiple $C_{1-\alpha}$'s that satisfy (5.1). For instance, given a posterior credibility region of ϕ with credibility $1-\alpha$, $B_{1-\alpha} \subset \Phi$, $C_{1-\alpha} = \cup_{\phi \in B_{1-\alpha}} H(\phi)$ considered in the 2009 working paper version of Moon and Schorfheide (2011) satisfies (5.1).

In proposing set inference for η, I resolve the multiplicity issue of set estimates by focusing on the smallest one,

$$\begin{aligned} & C_{1-\alpha*} \equiv \arg\min_{C \in \mathcal{C}} Leb(C) \\ & \text{s.t.} \quad F_{\phi|X}(H(\phi) \subset C)) \geq 1-\alpha, \end{aligned} \tag{5.2}$$

where $Leb(C)$ is the volume of subset C in terms of the Lebesgue measure and $\mathcal{C}$ is a family of subsets in $\mathcal{H}$ over which the volume-minimizing lower credible region is searched. I thereafter refer to $C_{1-\alpha*}$ defined in this way as a *posterior lower credible region* with credibility $1-\alpha$. Note that focusing on the smallest set estimate has a decision theoretic justification; $C_{1-\alpha*}$ can be supported as a posterior gamma minimax action:

$$C_{1-\alpha*} = \arg\min_{C \in \mathcal{C}} \left[\sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \int L(\eta, C)\, dF_{\eta|x} \right]$$

with a loss function that penalizes the volume and non-coverage,

$$L(\eta, C) = Leb(C) + b(\alpha)\left[1 - 1_C(\eta)\right],$$

where $b(\alpha)$ is a positive constant that depends on credibility level $1-\alpha$, and $1_C(\cdot)$ is the indicator function for subset C. Here, the loss function is written in terms of parameter η, so the object of interest is η, rather than identified set $H(\phi)$.

Finding $C_{1-\alpha*}$ is challenging if η is multi-dimensional and no restriction is placed on class of subsets $\mathcal{C}$. I therefore restrict our analysis to scalar η and constrain $\mathcal{C}$ to the class of closed connected intervals. The next proposition shows how to obtain $C_{1-\alpha*}$.

Proposition 5.1 *Let $d : \mathcal{H} \times \mathcal{D} \to \mathcal{R}_+$ measures distance from $\eta_c \in \mathcal{H}$ to set $H(\phi)$ in terms of*

$$\bar{d}(\eta_c, H(\phi)) \equiv \sup_{\eta \in H(\phi)} \{\|\eta_c - \eta\|\}.$$

For each $\eta_c \in \mathcal{H}$, let $r_{1-\alpha}(\eta_c)$ be the $(1-\alpha)$-th quantile of the distribution of $\bar{d}(\eta_c, H(\phi))$ induced by the posterior distribution of ϕ, i.e.,

$$r_{1-\alpha}(\eta_c) \equiv \inf \left\{ r : F_{\phi|X}\left(\left\{ \phi : \bar{d}(\eta_c, H(\phi)) \leq r \right\} \right) \geq 1 - \alpha \right\}.$$

Then, $C_{1-\alpha*}$ is a closed interval centered at $\eta_c^* = \arg\min_{\eta_c \in \mathcal{H}} r_{1-\alpha}(\eta_c)$ with radius $r_{1-\alpha}^* = r_{1-\alpha}(\eta_c^*)$.

Proof. See Appendix A. ■

5.2 Asymptotic Properties of the Posterior Lower Credible Region

In this section, I examine the large-sample behavior of the posterior lower probability in an intervally identified case, where $H(\phi) \subset \mathcal{R}$ is a closed bounded interval for almost all $\phi \in \Phi$. From now on, I make the sample size explicit in the notation: a size n sample X^n is generated from its sampling distribution $P_{X^n|\phi_0}$, where ϕ_0 denotes the value of the sufficient parameters that corresponds to the true data-generating process. The maximum likelihood estimator for ϕ is denoted by $\hat{\phi}$.

Provided that the posterior of ϕ is consistent[11] to ϕ_0 and the set-valued map $H(\phi_0)$ is continuous at $\phi = \phi_0$ in terms of the Hausdorff metric d_H, it can be shown that random sets $H(\phi)$, represented by the posterior lower probability $F_{\eta|X^n}(\cdot)$, converges to true identified set $H(\phi_0)$ in the sense of $\lim_{n \to \infty} F_{\phi|X^n}(\{\phi : d_H(H(\phi), H(\phi_0)) > \epsilon\}) = 0$ for almost

[11] Posterior consistency of ϕ means that $\lim_{n\to\infty} F_{\phi|X^n}(G) = 1$ for every G open neighborhood of ϕ_0 for almost every sampling sequence. For finite dimensional ϕ, this posterior consistency for ϕ is implied by a set of higher-level conditions for the likelihood of ϕ. We do not list up all those conditions here for the sake of brevity. See Section 7.4 of Schervish (1995) for details.

every sampling sequences of $\{X^n\}$. Given such posterior consistency of ϕ, we analyze the asymptotic coverage property of the lower credible region $C_{1-\alpha*}$.

The following set of regularity conditions are imposed to show the asymptotic correct coverage property of $C_{1-\alpha*}$.

Condition 5.1

(i) The parameter of interest η is a scalar and the identified set $H(\phi)$ is μ_ϕ-almost surely a non-empty and connected interval, $H(\phi) = [\eta_l(\phi), \eta_u(\phi)]$, $-\infty \leq \eta_l(\phi) \leq \eta_u(\phi) \leq \infty$, and the true identified set $H(\phi_0) = [\eta_l(\phi_0), \eta_u(\phi_0)]$ is a bounded interval.

(ii) For sequence $a_n \to \infty$, random variables $\hat{L} = -a_n \left(\eta_l(\phi_0) - \eta_l(\hat{\phi})\right)$ and $\hat{U} = a_n \left(\eta_u(\phi_0) - \eta_u(\hat{\phi})\right)$ converges in distribution to bivariate random variables (L, U), whose cumulative distribution function $J(\cdot)$ on $\mathcal{R}^2$ is Lipschitz continuous and monotonically increasing in the sense of $J(c_l, c_u) < J(c_l + \epsilon, c_u + \epsilon)$ for any $\epsilon > 0$.

(iii) Define random variables $L^n(\phi) = -a_n \left(\eta_l(\phi) - \eta_l(\hat{\phi})\right)$ and $U^n(\phi) = a_n \left(\eta_u(\phi) - \eta_u(\hat{\phi})\right)$, whose distribution is induced by the posterior distribution of ϕ given sample X^n. The cumulative distribution function of $(L^n(\phi), U^n(\phi))$ given X^n denoted by $J^n(\cdot)$ is continuous almost surely under $\hat{p}(x^n|\phi_0)$ for all n.

(iv) At each $c \equiv (c_l, c_u) \in R^2$, the cumulative distribution function of $(L^n(\phi), U^n(\phi))$ given X^n, $J^n(c)$, converges in probability under $P_{X^n|\phi_0}$ to $J(c)$.

Conditions 5.1 (ii) and (iv) imply that the estimators for the lower and upper bounds of $H(\phi)$ attain the Bernstein–von Mises property: the sampling distribution of the bound estimators and the posterior distribution of the bounds coincide asymptotically in the sense of Theorem 7.101 in Schervish (1995). In case of finite dimensional ϕ, Condition 5.1 (ii) and (iv), with $a_n = \sqrt{n}$ and bivariate normal (L, U), are implied from the following set of assumptions: (a) the regularity of the likelihood of ϕ and the asymptotic normality of $\sqrt{n}\left(\phi - \hat{\phi}\right)$, (b) μ_ϕ puts a positive probability on every open neighborhood of ϕ_0 and μ_ϕ's density is smooth at ϕ_0, and (c) the applicability of the delta method to $\eta_l(\cdot)$ and $\eta_u(\cdot)$ at $\phi = \phi_0$ with non-zero first derivatives.[12]

[12] See Schervish (1995, Section 7.4) for further detail on these assumptions.

The next proposition establishes the large-sample coverage property of the posterior lower credible region $C_{1-\alpha *}$.

Theorem 5.1 (Asymptotic Coverage Property) *Assume Conditions 3.1 and 5.1.* $C_{1-\alpha *}$ *can be interpreted as frequentist confidence intervals for the true identified set* $H(\phi_0)$ *with a pointwise asymptotic coverage probability* $(1-\alpha)$,

$$\lim_{n\to\infty} P_{X^n|\phi_0}\left(H(\phi_0) \subset C_{1-\alpha *}\right) = 1-\alpha.$$

Proof. See Appendix A. ■

This result shows that, for the interval-identified η, the posterior lower credible region $C_{1-\alpha *}$ achieves the exact desired frequentist coverage for the *identified set* asymptotically (Horowitz and Manski (2000), Chernozhukov, Hong, and Tamer (2007), and Romano and Shaikh (2010)). It is worth noting that the posterior lower credible region $C_{1-\alpha *}$ differs from the confidence intervals for the *parameter of interest,* as considered in Imbens and Manski (2004) and Stoye (2009); in case $H(\phi_0)$ is an interval, $C_{1-\alpha *}$ will be asymptotically wider than the frequentist confidence interval for η. This implies that the set of priors $\mathcal{M}\left(\mu_\phi\right)$ is too large to interpret $C_{1-\alpha *}$ as the frequentist's confidence interval for η.

It is also worth noting that the asymptotic coverage probability presented in Theorem 5.1 is in the sense of pointwise asymptotic coverage rather than an asymptotic uniform coverage over ϕ_0. The frequentist literature has stressed the importance of the uniform coverage property of interval estimates in order to ensure that the intervals estimates can have an accurate coverage probability in a finite sample situation (Imbens and Manski (2004), Andrew and Guggenberger (2009), Stoye (2009), Romano and Shaikh (2010), Andrew and Soares (2010), among many others). Examining whether or not the posterior lower credible region constructed above can attain a uniformly valid coverage probability for the identified set is beyond the scope of this paper and is left for future research.

We note that Condition 5.1 (iv) is a quite delicate condition as illistrated by the following counterexample.

Example 5.1 *Let the identified set be given by* $H(\phi) = [\max\{\phi_1, \phi_2\}, \min\{\phi_3, \phi_4\}]$. *This type of bound commonly appears in the intersection bound analysis (Manski (1990)), and has* *attracted considerable attention in the literature (Hirano and Porter (2011), Chernozhukov, Lee, and Rosen (2011)). Condition 5.1 (iv) does not hold in this class of models when the true values of the arguments in the minimum or maximum happen to be equal.*

Let us focus on the lower bound $\eta_l(\phi) = \max\{\phi_1, \phi_2\}$. *Assume that the maximum likelihood estimators* $\hat{\phi}_1$ *and* $\hat{\phi}_2$ *are independent and* $\hat{\phi}_1 \sim N\left(\phi_{10}, \frac{1}{n}\right)$ *and* $\hat{\phi}_2 \sim N\left(\phi_{20}, \frac{1}{n}\right)$ *with* $\phi_{10} = \phi_{20}$. *As for the posterior of* ϕ_1 *and* ϕ_2*, assume that* $\phi_1|x^n \sim N\left(\hat{\phi}_1, \frac{1}{n}\right)$ *and* $\phi_2|x^n \sim N\left(\hat{\phi}_2, \frac{1}{n}\right)$. *In this case, the sampling distribution of* $L = -\sqrt{n}\left(\eta_l(\phi_0) - \eta_l(\hat{\phi})\right)$ *and the posterior distribution of* $L^n(\phi) = -\sqrt{n}\left(\eta_l(\phi) - \eta_l(\hat{\phi})\right)$ *are obtained as*

$$L \sim \max\left\{\begin{array}{c} Z_1 \\ Z_2 \end{array}\right\}, \qquad L^n(\phi)\,|x^n \sim \min\left\{\begin{array}{c} Z_1 + \left|\hat{Z}\right|_- \\ Z_2 + \left|\hat{Z}\right|_+ \end{array}\right\},$$

where (Z_1, Z_2) *are independent standard normal variables,* $\hat{Z} = \sqrt{n}\left(\hat{\phi}_1 - \hat{\phi}_2\right)$, $|a|_- = -\min\{0, a\}$*, and* $|a|_+ = \max\{0, a\}$. *The posterior distribution of* $L^n(\phi)$ *fails to converge to the sampling distribution of* $\sqrt{n}\left(\eta_l(\hat{\phi}) - \eta_l(\phi_0)\right)$ due to the non-vanishing $\hat{Z}$. *Note that, even when* $\hat{Z}$ *happens to be zero,* $L^n(\phi)$*'s posterior distribution differs from the sampling distribution of* L. $C_{1-\alpha*}$ *will estimate* $H(\phi_0)$ *with inward bias, and the coverage probability for* $H(\phi_0)$ *will be lower than the nominal coverage. A lesson from this example is that, despite the explicit introduction of ambiguity in the form of prior class* $\mathcal{M}(\mu_\phi)$ *and the decision theoretic justification behind the construction of* $C_{1-\alpha*}$*, the robust Bayes posterior inference procedure based on* $C_{1-\alpha*}$ *does not correct the frequentist bias issue in the intersection bounds analysis. In contrast, the correct coverage will be attained if* $C_{1-\alpha} = \cup_{\phi \in B_{1-\alpha}} H(\phi)$ *is used as a robustified posterior credible region.*

6 Concluding Remarks

This paper proposes a framework of a robust Bayes analysis for set-identified models in econometrics. I demonstrate that the posterior lower probability obtained from the prior class $\mathcal{M}(\mu_\phi)$ can be interpreted as the posterior probability law of the identified set (Theorem 3.1). This robust Bayesian way of generating and interpreting the identified set as an a posteriori random object has not been investigated in the literature, This highlights the seamless links among partial identification analysis, robust Bayes inference, and random set theory, and offers a unified framework of statistical decision and inference for set-identified parameters from the conditional perspective. I employ the posterior gamma-minimax criterion to formulate and solve for a statistical decision with multiple posteriors. The objective function of the gamma-minimax criterion integrates the ambiguity associated with the set identification and posterior uncertainty of the identified set into a single objective function. It leads to a numerically solvable posterior gamma-minimax action, as long as the identified sets $H(\phi)$ can be simulated from the posterior of ϕ.

The posterior lower probability is a non-additive measure, so one complication of the lower probability inference is that we cannot plot it as we would do for the posterior probability densities. To visualize it and conduct a set estimation in a decision-theoretically justifiable way, I propose the posterior lower credible region. For an interval-identified parameter, I derive the conditions that the posterior lower credible region with credibility $(1-\alpha)$ can be interpreted as an asymptotically valid frequentist confidence interval for the identified set with coverage $(1-\alpha)$. This claim can be seen as an extension of the celebrated Bernstein-von Mises theorem to the multiple prior Bayesian inference via the lower probability, and exemplifies a situation where the robust Bayesians can accomplish a compromise of the Bayesian and frequentist inference.

Appendix

A Lemmas and Proofs

In this appendix, I first demonstrate that the set-valued mappings $\Gamma(\phi)$ and $H(\phi)$ defined in the main text are closed random sets (measurable and closed set-valued mappings) induced by a probability measure on $(\Phi, \mathcal{B})$.

Lemma A.1 *Assume* $(\Theta, \mathcal{A})$ *and* $(\Phi, \mathcal{B})$ *are complete separable metric spaces. Under Condition 3.1,* $\Gamma(\phi)$ *and* $H(\phi)$ *are random closed sets induced by a probability measure on* $(\Phi, \mathcal{B})$*,* *i.e.*, $\Gamma(\phi)$ *and* $H(\phi)$ *are closed and, for* $A \in \mathcal{A}$ *and* $D \in \mathcal{H}$,

$$\begin{aligned} \{\phi : \Gamma(\phi) \cap A \neq \emptyset\} &\in \mathcal{B} \quad \textit{for} \quad A \in \mathcal{A}, \\ \{\phi : H(\phi) \cap D \neq \emptyset\} &\in \mathcal{B} \quad \textit{for} \quad D \in \mathcal{H}. \end{aligned}$$

Proof. Closedness of $\Gamma(\phi)$ and $H(\phi)$ is implied directly from Conditions 3.1 (ii) and (iii). To prove the measurability of $\{\phi : \Gamma(\phi) \cap A \neq \emptyset\}$, Theorem 2.6 in Molchanov is invoked, which states that, given $(\Theta, \mathcal{A})$ as Polish, $\{\phi : \Gamma(\phi) \cap A \neq \emptyset\} \in \mathcal{B}$ holds if and only if $\{\phi : \theta \in \Gamma(\phi)\} \in \mathcal{B}$ is true for every $\theta \in \Theta$. Since $\Gamma(\phi)$ is an inverse image of the many-to-one and onto mapping, $g : \Theta \to \Phi$, a unique value of $\phi \in \Phi$ exists for each $\theta \in \Theta$, and $\{\phi\} \in \mathcal{B}$, since Φ is a metric space. Hence, $\{\phi : \theta \in \Gamma(\phi)\} \in \mathcal{B}$ holds.

To verify the measurability of $\{\phi : H(\phi) \cap D \neq \emptyset\}$, note that

$$\{\phi : H(\phi) \cap D \neq \emptyset\} = \left\{\phi : \Gamma(\phi) \cap h^{-1}(D) \neq \emptyset\right\}.$$

Since $h^{-1}(D) \in \mathcal{A}$, by the measurability of h (Condition 3.1 (iii)), the first statement of this lemma implies $\{\phi : H(\phi) \cap D \neq \emptyset\} \in \mathcal{B}$. ■

A.1 Proof of Theorem 3.1

Given the measurability $\Gamma(\phi)$ and $H(\phi)$, as proven in Lemma A.1, the proof of Theorem 3.1 given below uses the following two lemmas. The first lemma says that, given a fixed subset $A \in \mathcal{A}$ in the parameter space of θ, the conditional probability $\mu_{\theta|\phi}(A|\phi)$ can be bounded below by the indicator function $1_{\{\Gamma(\phi) \subset A\}}(\phi)$ when $\mu_\theta \in \mathcal{M}(\mu_\phi)$. The second lemma shows that for each fixed subset $A \in \mathcal{A}$, we can construct a probability measure on $(\Theta, \mathcal{A})$ that belongs to the prior class $\mathcal{M}(\mu_\phi)$ and achieves the lower bound of the conditional probability obtained in the first lemma. Theorem 3.1 follows as a corollary of these two lemma.

Lemma A.2 *Assume Condition 3.1 and let* $A \in \mathcal{A}$ *be an arbitrary fixed subset of* Θ. *For every* $\mu_\theta \in \mathcal{M}(\mu_\phi)$,

$$1_{\{\Gamma(\phi) \subset A\}}(\phi) \leq \mu_{\theta|\phi}(A|\phi)$$

holds μ_ϕ*-almost surely.*

Proof. For the given subset A, define $\Phi_1^A = \{\phi : \Gamma(\phi) \subset A\} \subset \Phi$. Note that, by Lemma A.1, Φ_1^A belongs to the sufficient parameter σ-algebra $\mathcal{B}$. To prove the claim, it suffices to show that

$$\int_B 1_{\Phi_1^A}(\phi) d\mu_\phi \leq \int_B \mu_{\theta|\phi}(A|\phi) d\mu_\phi \tag{A.1}$$

for every $\mu_\theta \in \mathcal{M}(\mu_\phi)$ and $B \in \mathcal{B}$.

Consider

$$\int_B \mu_{\theta|\phi}(A|\phi) d\mu_\phi \geq \int_{B \cap \Phi_1^A} \mu_{\theta|\phi}(A|\phi) d\mu_\phi = \mu_\theta(A \cap \Gamma(B \cap \Phi_1^A)).$$

where the equality follows by the definition of the conditional probability. By the construction of Φ_1^A, $\Gamma(B \cap \Phi_1^A) \subset A$ holds, so

$$\begin{aligned} \mu_\theta(A \cap \Gamma(B \cap \Phi_1^A)) &= \mu_\theta(\Gamma(B \cap \Phi_1^A)) \\ &= \mu_\phi(B \cap \Phi_1^A) \\ &= \int_B 1_{\Phi_1^A}(\phi) d\mu_\phi. \end{aligned}$$

Thus, the inequality (A.1) is proven. ■

Lemma A.3 *Assume Condition 3.1. For each $A \in \mathcal{A}$, there exists $\mu_{\theta*} \in \mathcal{M}(\mu_\phi)$ whose conditional distribution $\mu_{\theta*|\phi}$ achieves the lower bound of $\mu_{\theta|\phi}(A|\phi)$ obtained in Lemma A.2, μ_ϕ-almost surely.*

Proof. Fix subset $A \in \mathcal{A}$ throughout the proof. Consider partitioning the sufficient parameter space Φ into three, based on the relationship between $\Gamma(\phi)$ and A,

$$\begin{aligned} \Phi_0^A &= \{\phi : \Gamma(\phi) \cap A = \emptyset\}, \\ \Phi_1^A &= \{\phi : \Gamma(\phi) \subset A\}, \\ \Phi_2^A &= \{\phi : \Gamma(\phi) \cap A \neq \emptyset \text{ and } \Gamma(\phi) \cap A^c \neq \emptyset\}, \end{aligned}$$

where each of Φ_0^A, Φ_1^A, and Φ_2^A belongs to the sufficient parameter σ-algebra $\mathcal{B}$ by Lemma A.1. Note that Φ_0^A, Φ_1^A, and Φ_2^A are mutually disjoint and constitute a partition of Φ.

Now, consider a Θ-valued measurable selection $\xi^A(\cdot)$ defined on Φ_2^A such that $\xi^A(\phi) \in [\Gamma(\phi) \cap A^c]$ holds for μ_ϕ-almost every $\phi \in \Phi_2^A$. Note that such measurable selection

$\xi^A(\phi)$ can be constructed, for instance, by $\xi^A(\phi) = \arg\max_{\theta \in \Gamma(\phi)} d(\theta, A)$, where $d(\theta, A) = \inf_{\theta' \in A} \|\theta - \theta'\|$ (see Theorem 2.27 in Chapter 1 of Molchanov (2005) for B-measurability of such $\xi^A(\phi)$). Let us pick a probability measure from the prior class, $\mu_\theta \in \mathcal{M}(\mu_\phi)$, and construct another measure $\mu_{\theta*}$ by

$$\mu_{\theta*}\left(\tilde{A}\right) = \mu_\theta\left(\tilde{A} \cap \Gamma\left(\Phi_0^A\right)\right) + \mu_\theta\left(\tilde{A} \cap \Gamma\left(\Phi_1^A\right)\right) + \mu_\phi\left(\left\{\xi^A(\phi) \in \tilde{A}\right\} \cap \Phi_2^A\right), \qquad \tilde{A} \in \mathcal{A}.$$

It can be checked that $\mu_{\theta*}$ is a probability measure on $(\Theta, \mathcal{A})$: $\mu_{\theta*}$ satisfies $\mu_{\theta*}(\emptyset) = 0$, $\mu_{\theta*}(\Theta) = 1$, and countable additivity. Furthermore, $\mu_{\theta*}$ belongs to $\mathcal{M}(\mu_\phi)$ because, for $B \in \mathcal{B}$,

$$\begin{aligned}
\mu_{\theta*}(\Gamma(B)) &= \mu_\theta\left(\Gamma(B) \cap \Gamma\left(\Phi_0^A\right)\right) + \mu_\theta\left(\Gamma(B) \cap \Gamma\left(\Phi_1^A\right)\right) + \mu_\phi\left(\left\{\xi^A(\phi) \in \Gamma(B)\right\} \cap \Phi_2^A\right) \\
&= \mu_\theta\left(\Gamma\left(B \cap \Phi_0^A\right)\right) + \mu_\theta\left(\Gamma\left(B \cap \Phi_1^A\right)\right) + \mu_\phi\left(B \cap \Phi_2^A\right) \\
&= \mu_\phi\left(B \cap \Phi_0^A\right) + \mu_\phi\left(B \cap \Phi_1^A\right) + \mu_\phi\left(B \cap \Phi_2^A\right) \\
&= \mu_\phi(B),
\end{aligned}$$

where the second line follows because $\Gamma(\phi)$'s are disjoint and $\xi(\phi) \in \Gamma(\phi)$ holds for almost every $\phi \in \Phi_2^A$. With the thus-constructed $\mu_{\theta*}$ and an arbitrary subset $B \in \mathcal{B}$, consider

$$\begin{aligned}
\mu_{\theta*}(A \cap \Gamma(B)) &= \mu_\theta(A \cap \Gamma(B) \cap \Gamma(\Phi_0^A)) + \mu_\theta(A \cap \Gamma(B) \cap \Gamma(\Phi_1^A)) \\
&\quad + \mu_\phi\left(\left\{\xi^A(\phi) \in [A \cap \Gamma(B)]\right\} \cap \Phi_2^A\right).
\end{aligned}$$

Here, by the construction of $\left\{\Phi_j^A\right\}_{j=1,2,3}$ and $\xi^A(\phi)$, it holds that $A \cap \Gamma(\Phi_0^A) = \emptyset$, $\Gamma(\Phi_1^A) \subset A$, and $\mu_\phi\left(\left\{\xi^A(\phi) \in [A \cap \Gamma(B)]\right\} \cap \Phi_2^A\right) = 0$. Accordingly, we obtain

$$\begin{aligned}
\mu_{\theta*}(A \cap \Gamma(B)) &= \mu_\theta(\Gamma(B) \cap \Gamma(\Phi_1^A)) \\
&= \mu_\phi\left(B \cap \Phi_1^A\right) \\
&= \int_B 1_{\Phi_1^A}(\phi) d\mu_\phi.
\end{aligned}$$

Since $B \in \mathcal{B}$ is arbitrary, this implies that $\mu_{\theta*}(A|\phi) = 1_{\Phi_1^A}(\phi)$, μ_ϕ-almost surely. Thus, $\mu_{\theta*}$ achieves the lower bound obtained in Lemma A.2. ■

Proof of Theorem 3.1 (i). Under the given assumptions, the posterior of θ is given by (see equation (3.4))

$$F_{\theta|X}(A) = \int_\Phi \mu_{\theta|\phi}(A|\phi) dF_{\phi|X}(\phi).$$

By the monotonicity of the integral, $F_{\theta|X}(A)$ is minimized over the prior class by plugging in the attainable pointwise lower bound of $\mu_{\theta|\phi}(A|\phi)$ into the integrand. By Lemmas A.2 and A.3, the attainable pointwise lower bound of $\mu_{\theta|\phi}(A|\phi)$ is given by $1_{\{\Gamma(\phi)\subset A\}}(\phi)$, so that it holds

$$F_{\theta|X*}(A) = \int_{\Phi} 1_{\{\Gamma(\phi)\subset A\}}(\phi)dF_{\phi|X}(\phi) = F_{\phi|X}(\{\phi : \Gamma(\phi) \subset A\}).$$

The posterior upper probability follows by its conjugacy with the lower probability:

$$F^*_{\theta|X}(A) = 1 - F_{\theta|X*}(A^c) = F_{\phi|X}(\{\phi : \Gamma(\phi) \cap A \neq \emptyset\}).$$

■

A.2 Proof of Proposition 4.1

The next lemma is used to prove Proposition 4.1. This lemma states that the core of $F_{\theta|X*}$ agrees with the set of posteriors of θ induced by prior class $\mathcal{M}(\mu_\phi)$. In the terminology of Huber (1973), this property is called the representability of the class of probability measures by the lower and upper probabilities.

Lemma A.4 *Assume Condition 3.1. Define the core of the posterior lower probability,*

$$core\left(F_{\theta|X*}\right) = \left\{G_\theta : G_\theta \textit{ probability measure on } (\Theta, \mathcal{A})\,,\, F_{\theta|X*}(A) \leq G_\theta(A) \textit{ for every } A \in \mathcal{A}\right\}.$$

It holds that

$$core\left(F_{\theta|X*}\right) = \left\{F_{\theta|X} : F_{\theta|X} \textit{ posterior distribution on } (\Theta, \mathcal{A}) \textit{ induced by some } \mu_\theta \in \mathcal{M}(\mu_\phi)\right\}.$$

Proof of Lemma A.4. For each $\mu_\theta \in \mathcal{M}(\mu_\phi)$, $F_{\theta|X*}(A) \leq F_{\theta|X}(A) \leq F^*_{\theta|X}(A)$ holds for every $A \in \mathcal{A}$, by the definition of the lower and upper probabilities. Hence, $core\left(F_{\theta|X*}\right)$ contains $\left\{F_{\theta|X} : \mu_\theta \in \mathcal{M}(\mu_\phi)\right\}$.

To show the converse, recall Theorem 3.1 (i), which shows that the lower and upper probabilities are the containment and capacity functionals of the random closed set $\Gamma(\phi)$.

As a result, by applying the selectionability theorem of random sets (Molchanov (2005), Theorem 1.2.20), it holds that for each $G_\theta \in core\left(F_{\theta|X*}\right)$, there exists a Θ-valued random variable $\xi(\phi)$, a so-called measurable selection of $\Gamma(\phi)$, such that $\xi(\phi) \in \Gamma(\phi)$ holds for every $\phi \in \Phi$ and $G_\theta(A) = F_{\phi|X}\left(\xi(\phi) \in A\right)$, $A \in \mathcal{A}$. Let $G_\theta \in core\left(F_{\theta|X*}\right)$ be fixed and let μ_θ^ξ be the probability distribution of a corresponding measurable selection $\xi\left(\phi\right)$ induced by the prior of ϕ:

$$\mu_\theta^\xi(A) = \mu_\phi\left(\{\phi : \xi(\phi) \in A\}\right).$$

Note that such a μ_θ^ξ belongs to $\mathcal{M}(\mu_\phi)$ since, for each subset $B \in \mathcal{B}$ in the sufficient parameter space,

$$\mu_\theta^\xi(\Gamma(B)) = \mu_\phi\left(\{\phi : \xi(\phi) \in \Gamma(B)\}\right) = \mu_\phi(B),$$

where the second equality holds because $\{\Gamma(\phi) : \phi \in B\}$ are mutually disjoint and $\xi(\phi) \in \Gamma(\phi)$ for every ϕ. Since the conditional distribution of $\mu_\theta^\xi(A)$ given ϕ is $\mu_{\theta|\phi}^\xi(A|\phi) = 1_{\{\xi(\phi)\in A\}}(\phi)$, the posterior distribution of θ generated from μ_θ^ξ is, by (3.4),

$$\begin{aligned}
\tilde{F}_{\theta|X}(A) &= \int 1_{\{\xi(\phi)\in A\}}(\phi) dF_{\phi|X}(\phi) \\
&= F_{\phi|X}\left(\xi(\phi) \in A\right) \\
&= G_\theta(A).
\end{aligned}$$

Thus, it is shown that for each $G_\theta \in \mathcal{G}_\theta$, there exists a prior $\mu_\theta^\xi \in \mathcal{M}(\mu_\phi)$, with which the posterior of θ coincides with G_θ. Hence, $core\left(F_{\theta|X*}\right) \subset \left\{F_{\theta|X} : \mu_\theta \in \mathcal{M}(\mu_\phi)\right\}$. ∎

Proof of Proposition 4.1. Let $core\left(F_{\theta|X*}\right)$ be the core of $F_{\theta|X*}$ as defined in Lemma A.4. I apply Proposition 10.3 in Denneberg (1994a), which states that if the upper probability $F_{\theta|X}^*(\cdot)$ is submodular, then, for any non-negative measurable function $k : \Theta \to \mathcal{R}_+$,

$$\int k(\theta) dF_{\theta|X}^* = \sup_{G_\theta \in core\left(F_{\theta|X*}\right)} \left\{ \int_\Theta k(\theta) dG_\theta \right\} \tag{A.2}$$

holds. Corollary 3.1 assures submodularity of $F_{\theta|X}^*(\cdot)$, and Lemma A.4 implies that $\sup_{G_\theta \in core\left(F_{\theta|X*}\right)} \left\{ \int_\Theta k(\theta\right.$ is equivalent to $\sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \left\{ \int_\Theta k(\theta) dF_{\theta|X} \right\}$. Hence, setting $k(\theta) = L(h(\theta), a)$ in (A.2) leads to the equality of the Choquet integral of $L(h(\theta), a)$ with respect to $F^*_{\theta|X}$ to the posterior upper risk:

$$\begin{aligned}
\int L(h(\theta), a) dF^*_{\theta|X} &= \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \left\{ \int_\Theta L(h(\theta), a) dF_{\theta|X} \right\} \\
&= \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \left\{ \int_{\mathcal{H}} L(\eta, a) dF_{\eta|X} \right\} \\
&= \rho^* \left(\mu_\phi, a \right).
\end{aligned} \tag{A.3}$$

Note further that, by the definition of the Choquet integral and Theorem 3.1 (ii):

$$\begin{aligned}
\int L(\eta, a) dF^*_{\eta|X} &= \int F^*_{\eta|X}(\{\eta : L(\eta, a) \geq t\}) dt \\
&= \int F^*_{\theta|X}\left(\{\theta : L(h(\theta), a)\}\right) dt \\
&= \int L(h(\theta), a) dF^*_{\theta|X}.
\end{aligned} \tag{A.4}$$

Combining (A.3) and (A.4) yields the first equality of the proposition.

The second equality of the proposition follows from Theorem 5.1 in Molchanov (2005).

■

A.3 Proof of Proposition 5.1

Proof of Proposition 5.1. The event $\{H(\phi) \subset C_r(\eta_c)\}$ happens if and only if $\left\{\bar{d}(\eta_c, H(\phi)) \leq r\right\}$. So, $r_{1-\alpha}(\eta_c) \equiv \inf \left\{ r : F_{\phi|X}\left(\left\{ \phi : \bar{d}(\eta_c, H(\phi)) \leq r \right\} \right) \geq 1 - \alpha \right\}$ is the radius of the smallest interval centered at η_c that contains random sets $H(\phi)$ with the posterior probability of at least $(1 - \alpha)$. Therefore, finding a minimizer of $r_{1-\alpha}(\eta_c)$ in η_c is equivalent to searching for the center of the smallest interval that contains $H(\phi)$ with posterior probability $1 - \alpha$. The attained minimum of $r_{1-\alpha}(\eta_c)$ provides its radius. ■

A.4 Proof of Theorem 5.1

Throughout the proofs, superscript n will be used to denote random objects induced by size n sample $X^n \sim P_{X^n|\phi_0}$, the probability law of X^n at $\phi = \phi_0$. Convergence in probability under $P_{X^n|\phi_0}$ is denoted by "$\underset{P}{\longrightarrow}$". I abbreviate "holds for infinitely many n" by "i.o." (infinitely often).

The following three lemmas will be used.

Lemma A.5 *Let $J^n(\cdot)$ and $J(\cdot)$ be as defined in Condition 5.1. Under Condition 5.1,*

$$\sup_{c \in \mathcal{R}^2} |J^n(c) - J(c)| \underset{P}{\longrightarrow} 0.$$

Proof. Lemma 2.11 in van der Vaart (1998) shows a non-stochastic version of this lemma. It is straightforward to extend it to the case where the pointwise convergence of J^n to J is assumed in terms of convergence in probability with respect to $P_{X^n|\phi_0}$. ■

Lemma A.6 *Let $Lev^n_{1-\alpha}$ and $Lev_{1-\alpha}$ be the $(1-\alpha)$-level sets of $J^n(\cdot)$ and $J(\cdot)$,*

$$\begin{aligned} Lev^n_{1-\alpha} &= \left\{c \in \mathcal{R}^2 : J^n(c) \geq 1-\alpha\right\}, \\ Lev_{1-\alpha} &= \left\{c \in \mathcal{R}^2 : J(c) \geq 1-\alpha\right\}. \end{aligned}$$

Define a distance from point $c \in \mathcal{R}^2$ to set $C \subset \mathcal{R}^2$ in terms of $d(c, C) \equiv \inf_{c' \in C} \|c - c'\|$, where $\|\cdot\|$ is the Eucledean distance. Under Condition 5.1, (a) $d\left(c, Lev^n_{1-\alpha}\right) \underset{P}{\longrightarrow} 0$ for every $c \in Lev_{1-\alpha}$, and (b) $d(c^n, Lev_{1-\alpha}) \underset{P}{\longrightarrow} 0$ for every $\{c^n : n = 1, 2, \dots\}$ sequence of measurable selections of $Lev^n_{1-\alpha}$.

Proof. To prove (a), suppose the conclusion is false. That is, there exist ϵ, $\delta > 0$, and $c = (c_l, c_u) \in Lev_{1-\alpha}$ such that $P_{X^n|\phi_0}\left(d\left(c, Lev^n_{1-\alpha}\right) > \epsilon\right) > \delta$, i.o. Event $d\left(c, Lev^n_{1-\alpha}\right) > \epsilon$ implies $J^n\left(c_l + \frac{\epsilon}{2}, c_u + \frac{\epsilon}{2}\right) < 1 - \alpha$ since $\left(c_l + \frac{\epsilon}{2}, c_u + \frac{\epsilon}{2}\right) \notin Lev^n_{1-\alpha}$. Therefore, it holds that

$$P_{X^n|\phi_0}\left(J^n\left(c_l + \frac{\epsilon}{2}, c_u + \frac{\epsilon}{2}\right) < 1 - \alpha\right) > \delta, \text{ i.o.} \tag{A.5}$$

Under Condition 5.1 (iv), however,

$$J^n\left(c_l + \frac{\epsilon}{2}, c_u + \frac{\epsilon}{2}\right) - J\left(c_l + \frac{\epsilon}{2}, c_u + \frac{\epsilon}{2}\right) \underset{P}{\longrightarrow} 0.$$

This convergence combined with

$$J\left(c_l + \frac{\epsilon}{2}, c_u + \frac{\epsilon}{2}\right) > J(c) \geq 1 - \alpha$$

due to strict monotonicity of $J(\cdot)$ implies that $P_{X^n|\phi_0}\left(J^n\left(c_l+\frac{\epsilon}{2}, c_u+\frac{\epsilon}{2}\right) \geq 1-\alpha\right) \to 1$ as $n \to \infty$. This contradicts (A.5).

To prove (b), suppose again the conclusion is false. This implies there exist ϵ, $\delta > 0$, and a sequence of random variables (measurable selections), $c^n = (c_l^n, c_u^n)$ with $P_{X^n|\phi_0}\left(c^n \in Lev_{1-\alpha}^n\right) = 1$ for all n, such that $P_{X^n|\phi_0}\left(d\left(c^n, Lev_{1-\alpha}\right) > \epsilon\right) > \delta$, i.o. Event $d\left(c^n, Lev_{1-\alpha}\right) > \epsilon$ implies $J\left(c_l^n+\frac{\epsilon}{2}, c_u^n+\frac{\epsilon}{2}\right) < 1-\alpha$, so it holds that

$$P_{X^n|\phi_0}\left(J\left(c_l^n+\frac{\epsilon}{2}, c_u^n+\frac{\epsilon}{2}\right) < 1-\alpha\right) > \delta, \text{ i.o.} \tag{A.6}$$

To find contradiction, note that

$$\begin{aligned} J\left(c_l^n+\frac{\epsilon}{2}, c_u^n+\frac{\epsilon}{2}\right) &= \left[J\left(c_l^n+\frac{\epsilon}{2}, c_u^n+\frac{\epsilon}{2}\right) - J\left(c_l^n, c_u^n\right)\right] \\ &\quad + \left[J\left(c_l^n, c_u^n\right) - J^n\left(c_l^n, c_u^n\right)\right] + J^n\left(c_l^n, c_u^n\right) \\ &> \left[J\left(c_l^n, c_u^n\right) - J^n\left(c_l^n, c_u^n\right)\right] + 1-\alpha \\ &\xrightarrow[P]{} 1-\alpha, \end{aligned}$$

where the convergence in probability in the last line follows from $\left[J\left(c_l^n, c_u^n\right) - J^n\left(c_l^n, c_u^n\right)\right] \to_P 0$ as implied by Lemma A.5. This implies $P_{X^n|\phi_0}\left(J\left(c_l^n+\frac{\epsilon}{2}, c_u^n+\frac{\epsilon}{2}\right) \geq 1-\alpha\right) \to 1$ as $n \to \infty$, which contradicts (A.6). ■

Lemma A.7 *Assume Condition 5.1. Let*

$$\begin{aligned} K^n &= \arg\min\left\{c_l+c_u : J^n(c) \geq 1-\alpha\right\}, \\ K &= \arg\min\left\{c_l+c_u : J(c) \geq 1-\alpha\right\}, \end{aligned}$$

which are non-empty and closed since the objective function is continuous and monotonically decreasing, and level sets $Lev_{1-\alpha}^n$ and $Lev_{1-\alpha}$ are closed and bounded from below by continuity of J^n and J (as imposed in Conditions 5.1 (ii) and (iv)). Then, $d\left(\hat{c}^n, K\right) \xrightarrow[P]{} 0$ holds for every $\{\hat{c}^n : n = 1, 2, \dots\}$ sequence of measurable selection from K^n.

Proof. Suppose that the conclusion is false, that is, there exist ϵ, $\delta > 0$, and $\hat{c}^n = (\hat{c}_l^n, \hat{c}_u^n)$ with $P_{X^n|\phi_0}\left(\hat{c}^n \in K^n\right) = 1$ for all n, such that

$$P_{X^n|\phi_0}\left(d\left(\hat{c}^n, K\right) > \epsilon\right) > \delta, \quad \text{i.o.} \tag{A.7}$$

By the construction of $\hat{c}^n$, $P_{X^n|\phi_0}\left(\hat{c}^n \in Lev_{1-\alpha}^n\right) = 1$ for all n, so that Lemma A.6 (b) assures that there exists a random sequence, $\tilde{c}^n = (\tilde{c}_l^n, \tilde{c}_u^n)$ such that $P_{X^n|\phi_0}(\tilde{c}^n \in Lev_{1-\alpha}) = 1$ for all n and $\|\hat{c}^n - \tilde{c}^n\| \underset{P}{\longrightarrow} 0$ hold. Consequently, (A.7) implies that an analogous statement holds for $\tilde{c}^n$ as well; $P_{X^n|\phi_0}(d(\tilde{c}^n, K) > \epsilon) > \delta$, i.o. Let $\hat{f}^n = \hat{c}_l^n + \hat{c}_u^n$, $\tilde{f}^n = \tilde{c}_l^n + \tilde{c}_u^n$, and $f^* = \min\{c_l + c_u : J(c) \geq 1 - \alpha\}$. If $P_{X^n|\phi_0}(d(\tilde{c}^n, K) > \epsilon) > \delta$, i.o. is true, then $P_{X^n|\phi_0}(\tilde{c}^n \in Lev_{1-\alpha}) = 1$ implies that there exists $\xi > 0$ such that $P_{X^n|\phi_0}\left(\tilde{f}^n - f^* > \xi\right) > \delta$, i.o. Whereas, when $\left|\hat{f}^n - \tilde{f}^n\right| \underset{P}{\longrightarrow} 0$, it holds that

$$P_{X^n|\phi_0}\left(\hat{f}^n - f^* > \xi\right) > \delta, \quad \text{i.o.} \tag{A.8}$$

In order to derive a contradiction, pick $c \in K$ and apply Lemma A.6 (a) to find a random sequence $c^n = (c_l^n, c_u^n)$ such that $P_{X^n|\phi_0}\left(c^n \in Lev_{1-\alpha}^n\right) = 1$ for all n and $\|c - c^n\| \underset{P}{\longrightarrow} 0$. Along such sequence c^n, $f^* - (c_l^{n*} + c_u^{n*}) \underset{P}{\longrightarrow} 0$ holds. Then, (A.8) combined with $f^* - (c_l^n + c_u^n) \underset{P}{\longrightarrow} 0$ leads to

$$P_{X^n|\phi_0}\left(\hat{f}^n - (c_l^n + c_u^n) > \xi\right) > \delta, \ \text{i.o.},$$

implying that the value of the objective function evaluated at feasible point $c^n \in Lev_{1-\alpha}^n$ is smaller than that evaluated at $\hat{c}^n$ with a positive probability. This contradicts that $\hat{c}^n$ is a minimizer, $P_{X^n|\phi_0}(\hat{c}^n \in K^n) = 1$ for all n. ■

Proof of Theorem 5.1. By denoting a connected interval as $C = [l, u]$, I can write the optimization problem for obtaining $C_{1-\alpha*}$ as

$$\begin{aligned} & \min_{l,u} [u - l] \\ \text{s.t.} \quad & F_{\phi|X^n}(l \leq \eta_l(\phi) \ \text{ and } \ \eta_u(\phi) \leq u) \geq 1 - \alpha. \end{aligned}$$

In terms of random variables $L^n(\phi) = -a_n\left(\eta_l(\phi) - \eta_l(\hat{\phi})\right)$ and $U^n(\phi) = a_n\left(\eta_u(\phi) - \eta_u(\hat{\phi})\right)$, I can rewrite the constraint as:

$$F_{\phi|X^n}\left(L^n(\phi) \leq -a_n\left(l - \eta_l(\hat{\phi})\right) \ \text{ and } \ U^n(\phi) \leq a_n\left(u - \eta_u(\hat{\phi})\right)\right) \geq 1 - \alpha.$$

Therefore, by defining $c_l = -a_n\left(l - \eta_l(\hat{\phi})\right)$ and $c_u = a_n\left(u - \eta_u(\hat{\phi})\right)$, the above constrained minimization problem is written as

$$\begin{aligned} & \min\left[c_l + c_u\right] \\ \text{s.t.} \quad & J^n\left(c_l, c_u\right) \geq 1-\alpha. \end{aligned}$$

Let K^n be the set of solutions of this minimization problem, as defined in Lemma A.7. For a sequence of random variables $\hat{c}^n = (\hat{c}_l^n, \hat{c}_u^n)$ such that $P_{X^n|\phi_0}\left(\hat{c}^n \in K^n\right) = 1$, the posterior lower credible region is obtained as $C_{1-\alpha *} = \left[\eta_l(\hat{\phi}) - \frac{\hat{c}_l^n}{a_n}, \eta_u(\hat{\phi}) + \frac{\hat{c}_u^n}{a_n}\right]$. The coverage probability of $C_{1-\alpha *}$ for the true identified set is

$$\begin{aligned} & P_{X^n|\phi_0}\left(H\left(\phi_0\right) \subset C_{1-\alpha *}\right) \\ = \; & P_{X^n|\phi_0}\left(\eta_l\left(\hat{\phi}\right) - \frac{\hat{c}_l^n}{a_n} \leq \eta_l\left(\phi_0\right) \quad \text{and} \quad \eta_u\left(\phi_0\right) \leq \eta_u(\hat{\phi}) + \frac{\hat{c}_u^n}{a_n}\right) \\ = \; & P_{X^n|\phi_0}\left(-a_n\left(\eta_l\left(\phi_0\right) - \eta_l(\hat{\phi})\right) \leq \hat{c}_l^n \quad \text{and} \quad a_n\left(\eta_u\left(\phi_0\right) - \eta_u(\hat{\phi})\right) \leq \hat{c}_u^n\right) \\ \equiv \; & \hat{J}\left(\hat{c}^n\right), \end{aligned}$$

where $\hat{J}(\cdot)$ denotes the cumulative distribution function of the sampling distribution of $\left(-a_n\left(\eta_l\left(\phi_0\right) - \eta_l(\hat{\phi})\right),\ a_n\left(\eta_u\left(\phi_0\right) - \eta_u(\hat{\phi})\right)\right)$, which converges to $J(\cdot)$ pointwise by Condition 5.1 (ii). By Lemma A.7, there exists a random sequence $\bar{c}^n \in K$ such that $\|\bar{c}^n - \hat{c}^n\| \underset{P}{\longrightarrow} 0$. Hence, for $\epsilon > 0$ and such $\bar{c}^n$, it holds that

$$\begin{aligned} & P_{X^n|\phi_0}\left(H\left(\phi_0\right) \subset C_{1-\alpha *}\right) = \hat{J}\left(\hat{c}^n\right) \\ = \; & J\left(\bar{c}^n\right) + J\left(\hat{c}^n\right) - J\left(\bar{c}^n\right) + \hat{J}\left(\hat{c}^n\right) - J\left(\hat{c}^n\right) + o(1) \end{aligned} \tag{A.9}$$

Note that $J\left(\bar{c}^n\right) = 1-\alpha$ because, by strict monotonicity of the objective function and continuity of $J(\cdot)$, $\bar{c}^n \in K$ lies in the boundary of the level set $\{c \in \mathcal{R}^2 : J(c) \geq 1-\alpha\}$. By the Lipschitz continuity of $J(\cdot)$ (Condition 5.1 (ii)), for positive constant $M < \infty$,

$$\left|J\left(\hat{c}^n\right) - J\left(\bar{c}^n\right)\right| \leq M\epsilon P_{X^n|\phi_0}\left(\|\hat{c}^n - \bar{c}^n\| \leq \epsilon\right) + 2P_{X^n|\phi_0}\left(\|\hat{c}^n - \bar{c}^n\| > \epsilon\right) = M\epsilon + o(1).$$

Furthermore, $\left|\hat{J}\left(\hat{c}^n\right) - J\left(\hat{c}^n\right)\right| = o(1)$ holds by the uniform convergence of $\hat{J}(\cdot)$ to $J(\cdot)$ (Lemma 2.11 in van der Vaart (1998)). Therefore, (A.9) leads to

$$\left|P_{X^n|\phi_0}\left(H\left(\phi_0\right) \subset C_{1-\alpha *}\right) - (1-\alpha)\right| \leq M\epsilon + o(1).$$

Since $\epsilon > 0$ is arbitrary, I complete the proof. ■

B More on the Gamma-minimax Decision Analysis

B.1 Gamma-minimax Decision and Dynamic Inconsistency

This appendix examines the *unconditional* gamma-minimax decision problem (Kudo (1967), Berger (1985, pp. 213–218), etc.). Let $\delta(\cdot)$ be a decision function that maps $x \in \mathbf{X}$ to the action space $\mathcal{H}_a \subset \mathcal{H}$ and let Δ be the space of decisions (a subset of measurable functions $\mathbf{X} \to \mathcal{H}_a$.) The Bayes risk is defined as:

$$r(\mu_\theta, \delta) = \int_\Theta \left[\int_\mathcal{X} L(h(\theta), \delta(x)) p(x|\theta) dx \right] d\mu_\theta. \tag{B.1}$$

Given the prior class $\mathcal{M}(\mu_\phi)$, the unconditional *gamma-minimax criterion* ranks decisions in terms of the upper Bayes risk, $r^*(\mu_\phi, \delta) \equiv \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} r(\mu_\theta, \delta)$. Accordingly, the optimal decision under this criterion is defined as follows.

Definition B.1 *An unconditional gamma-minimax decision $\delta^* \in \Delta$ is a decision rule that minimizes the upper Bayes risk,*

$$r^*(\mu_\phi, \delta^*) = \inf_{\delta \in \Delta} r^*(\mu_\phi, \delta) = \inf_{\delta \in \Delta} \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} r(\mu_\theta, \delta).$$

In the standard Bayes decision problem with a single prior, the Bayes rule that minimizes $r(\mu_\theta, \delta)$ coincides with the posterior Bayes action for every possible sample $x \in \mathbf{X}$. Therefore, being either unconditional or conditional on data does not matter for the actual action to be taken. With multiple priors, however, the decision rule that minimizes $r^*(\mu_\phi, \delta)$ does not, in general, coincide with the posterior gamma-minimax action (Betro and Ruggeri (1992)). The next example illustrates that such a dynamic inconsistency also arises with the current specification of prior class.

Example B.1 *Consider the identified set for $\eta \in \mathcal{R}$ to be given by $H(\phi) = [\phi, \phi + c]$ with a known constant $c > 0$. Data $X = (X_1, \ldots, X_n)$ are generated i.i.d. from $\mathcal{N}(\phi, 1)$. Let a prior for ϕ be $\mathcal{N}(0, 1)$. Then, the posterior of ϕ is normal with mean $\frac{n}{n+1}\bar{X}$ and variance $\frac{1}{n+1}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Specify the quadratic loss function, $L(\eta, a) = (\eta - a)^2$.*

Consider computing the posterior gamma minimax action. With the quadratic loss, the integrand in the third expression of Proposition 4.1 (4.2) is written as

$$\sup_{\eta\in H(\phi)} L(\eta,a) = \left[a-\left(\phi+\frac{c}{2}\right)\right]^2 + c\left|a-\left(\phi+\frac{c}{2}\right)\right| + \frac{c^2}{4}.$$

Therefore, the posterior upper risk for action a is given by

$$E_{\phi|X}\left[a-\left(\phi+\frac{c}{2}\right)\right]^2 + cE_{\phi|X}\left|a-\left(\phi+\frac{c}{2}\right)\right| + \frac{c^2}{4}.$$

Since the posterior of ϕ is symmetric around its posterior mean, the posterior mean of $\left(\phi+\frac{c}{2}\right)$ minimizes $E_{\phi|X}\left[a-\left(\phi+\frac{c}{2}\right)\right]^2$ and $E_{\phi|X}\left|a-\left(\phi+\frac{c}{2}\right)\right|$ simultaneously. Hence, the posterior gamma-minimax action is

$$a_X^* = E_{\phi|X^n}(\phi) + \frac{c}{2} = \frac{n}{n+1}\bar{X} + \frac{c}{2}.$$

Next, consider the unconditional gamma minimax criterion for a decision function $\delta(X)$. Note that the Bayes risk with a single prior is written as

$$\begin{aligned} r(\mu_\theta,\delta) &= \int_\Phi\int_{\mathcal{H}}\left\{Var_{X|\phi}(\delta(X)) + \left[E_{X|\phi}[\delta(X)]-\eta\right]^2\right\} d\mu_{\eta|\phi}d\mu_\phi \\ &= \int_\Phi Var_{X|\phi}(\delta(X))\,d\mu_\phi + \int_\Phi\int_{\mathcal{H}}\left[E_{X|\phi}[\delta(X)]-\eta\right]^2 d\mu_{\eta|\phi}d\mu_\phi, \end{aligned}$$

where $Var_{X|\phi}(\cdot)$ and $E_{X|\phi}(\cdot)$ are the variance and expectation with respect to the sampling distribution of X, which only depends on ϕ by definition. By the same argument as in obtaining the third expression of Proposition 4.1 (4.2), I obtain the upper Bayes risk over the prior class $\mathcal{M}(\mu_\phi)$ as:

$$\begin{aligned} r^*(\mu_\phi,\delta) &= \int_\Phi Var_{X|\phi}(\delta(X))\,d\mu_\phi + \int_\Phi \sup_{\eta\in H(\phi)}\left[E_{X|\phi}[\delta(X)]-\eta\right]^2 d\mu_\phi. \\ &= \int_\Phi Var_{X|\phi}(\delta(X))\,d\mu_\phi \\ &\quad + \int_\Phi\left\{\left[E_{X|\phi}[\delta(X)]-\left(\phi+\frac{c}{2}\right)\right]^2 + c\left|E_{X|\phi}[\delta(X)]-\left(\phi+\frac{c}{2}\right)\right| + \frac{c^2}{4}\right\}d\mu_\phi. \end{aligned}$$

The upper Bayes risk evaluated at the posterior gamma-minimax action, $\delta(X) = a_X^* = \frac{n}{n+1}\bar{X}+\frac{c}{2}$, is calculated as

$$r^*(\mu_\phi,a_X^*) = \frac{1+c(2/\pi)^{\frac{1}{2}}}{n+1} + \frac{c^2}{4}.$$

Consider now a decision rule $\tilde{\delta}(X) = \bar{X} + \frac{c}{2}$ *that is different from* a_X^*. *The upper Bayes risk of* $\tilde{\delta}(X)$ *is obtained as*

$$r^*(\mu_\phi, \tilde{\delta}) = \frac{1}{n} + \frac{c^2}{4},$$

which is strictly smaller than $r^*(\mu_\phi, a_X^*)$ *if* $cn\,(2/\pi)^{\frac{1}{2}} > 1$. *Therefore, for some* c *and* n, $\tilde{\delta}(X)$ *outperforms* a_X^* *in terms of the unconditional gamma minimax criterion. Hence, the posterior gamma-minimax action* a_X^* *fails to be optimal in the unconditional sense.*

With the specified prior class, a prior that supprts a_x^* to be a posterior Bayes action generally depends on realization of data x. Consequently, one cannot claim that decision $\delta(x) = a_x^*$ is supported as an unconditional Bayes decision with respect to any prior in the class. On the other hand, the saddle-point argument in the minimax problem shows that the unconditional gamma-minimax decision δ^* is ensured to be a Bayes decision with respect to a least favorable prior in the class. This difference, that δ^* is a Bayes decision while a_x^* is not, results in the descrepancy between the conditional and unconditional gamma minimax decisions.

To my knowledge, no consensus is available on whether one should condition on data or not in solving the gamma minimax problem. In the current context, however, the two decision criteria differ in terms of simplicity of computing the optimal actions. The posterior gamma minimax action is simpler to compute, as shown in the main text, whereas it does not seem to be the case for the unconditional gamma minimax decision.

B.2 Gamma-minimax Regret Analysis

In this appendix, I consider the posterior gamma-minimax *regret* criterion as an alternative to the posterior gamma-minimax criterion considered in Section 4. To maintain analytical tractability, I consider the point estimation problem for scalar η with quadratic loss, $L(\eta, a) = (\eta - a)^2$.

The statistical decision problem under the conditional gamma-minimax regret criterion is set up as follows.

Definition B.2 *Define the lower bound of the posterior risk, given μ_θ, by $\underline{\rho}(\mu_\theta) = \inf_{a \in \mathcal{H}} \rho(\mu_\theta, a)$. The posterior gamma-minimax regret action $a_x^{reg} \in \mathcal{H}$ solves*

$$\inf_{a \in \mathcal{H}_a} \sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \left\{ \rho(\mu_\theta, a) - \underline{\rho}(\mu_\theta) \right\}.$$

Since the loss function is quadratic, the posterior risk $\rho(\mu_\theta, a)$ for a given μ_θ is minimized at $\hat{\eta}_{\mu_\theta}$, the posterior mean of η. Therefore, the lower bound of the posterior risk is simply the posterior variance, $\underline{\rho}(\mu_\theta) = E_{\eta|X}((\eta - \hat{\eta}_{\mu_\theta})^2)$, and the posterior regret can be written as

$$\begin{aligned} \rho(\mu_\theta, a) - \underline{\rho}(\mu_\theta) &= E_{\eta|X}\left[(\eta - a)^2 - \left(\eta - \hat{\eta}_{\mu_\theta}\right)^2 \right] \\ &= E_{\eta|X}\left[(a - \hat{\eta}_{\mu_\theta})^2 \right]. \end{aligned}$$

Let $\left[\underline{\eta}_x, \bar{\eta}_x\right]$ be the range of the posterior mean of η when μ_θ varies over the prior class $\mathcal{M}(\mu_\phi)$, where $\left[\underline{\eta}_x, \bar{\eta}_x\right]$ is assumed to be bounded, $m\left(x|\mu_\phi\right)$-almost surely. Then, the posterior gamma-minimax regret is simplified to

$$\sup_{\mu_\theta \in \mathcal{M}(\mu_\phi)} \left\{ \rho(\mu_\theta, a) - \underline{\rho}(\mu_\theta) \right\} = \begin{cases} (\bar{\eta}_x - a)^2 & \text{for } a \le \frac{\underline{\eta}_x + \bar{\eta}_x}{2}, \\ \left(\underline{\eta}_x - a\right)^2 & \text{for } a > \frac{\underline{\eta}_x + \bar{\eta}_x}{2}. \end{cases} \tag{B.2}$$

Hence, the posterior gamma-minimax regret is minimized at $\frac{\underline{\eta}_x + \bar{\eta}_x}{2}$, yielding the posterior gamma minimax regret action as the midpoint decision, $a_x^{reg} = \frac{\underline{\eta}_x + \bar{\eta}_x}{2}$. Since $\underline{\eta}_x$ and $\bar{\eta}_x$ are seen as the posterior means of the lower and upper bounds of the identified sets, reporting a_x^{reg} is qualitatively similar to reporting the midpoint of the efficient estimators for the bounds, which is known to be the local asymptotic minimax regret decision obtained in Song (2012).

Since the lower bound of the posterior risk $\underline{\rho}(\mu_\theta)$, in general, depends on the prior $\mu_\theta \in \mathcal{M}\left(\mu_\phi\right)$, the posterior gamma-minimax regret action a_x^{reg} differs from the posterior gamma-minimax action a_x^* obtained in Section 4. For large samples, however, this difference disappears and a_x^{reg} and a_x^* converge to the midpoint of the true identified sets in case of a symmetric loss function.

References

[1] Andrews, D.W.K. and P. Guggenberger (2009) "Validity of Subsampling and 'Plug-in Asymptotics' Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669-709.

[2] Andrews, D.W.K. and G. Soares (2010) "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119-157.

[3] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.

[4] Balke, A. and J. Pearl (1997) "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171-1176.

[5] Barankin, E.W. (1960) "Sufficient Parameters: Solution of the Minimal Dimensionality Problem," *Annals of the Institute of Mathematical Statistics*, 12, 91-118.

[6] Beresteanu, A. and F. Molinari (2008) "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763-814.

[7] Beresteanu, A., I.S. Molchanov, and F. Molinari (2012) "Partial Identification Using Random Sets Theory," *Journal of Econometrics*, 166, 17-32.

[8] Berger, J.O. (1984) "The Robust Bayesian Viewpoint." In J. Kadane ed. *Robust of Bayesian Analysis*, Amsterdam: North-Holland

[9] Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis, Second edition.* Spring-Verlag.

[10] Berger, J.O. and L.M. Berliner (1986) "Robust Bayes and Empirical Bayes Analysis with ϵ-contaminated Priors," *Annals of Statistics*, 14, 461-486.

[11] Bertsimas, D. and J.N. Tsitsiklis (1997) *Introduction to Linear Optimization.* Athena Scientific, Nashua, NH.

[12] Betro, B. and Ruggeri, F. (1992) "Conditional Γ-minimax Actions under Convex Losses," *Communications in Statistics, Part A – Theory and Methods*, 21, 1051-1066.

[13] Chamberlain, G. and E.E. Leamer (1976) "Matrix Weighted Averages and Posterior Bounds," *Journal of the Royal Statistical Society, Series B*, 38, 73-84.

[14] Chamberlain, G. (2000) "Econometric Applications of Maxmin Expected Utility," Journal of Applied Econometrics, Vol. 15, 6, 625-644.

[15] Chernozhukov, V., H. Hong, and E. Tamer (2007) "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243-1284.

[16] Chernozhukov, V., S. Lee, and A. Rosen (2011) "Intersection Bounds: Estimation and Inference," *cemmap working paper.*

[17] Dawid, A.P. (1979) "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41, No. 1, 1-31.

[18] DeRobertis, L. and J.A. Hartigan (1981) "Bayesian Inference Using Intervals of Measures," *The Annals of Statistics*, 9, No. 2, 235-244.

[19] Dempster, A.P. (1966) "New Methods for Reasoning Towards Posterior Distributions Based on Sample Data," *Annals of Mathematical Statistics*, 37, 355-374.

[20] Dempster, A.P. (1967a) "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals of Mathematical Statistics*, 38, 325-339.

[21] Dempster, A.P. (1967b) "Upper and Lower Probability Inferences Based on a Sample from a Finite Univariate Population," *Biometrika*, 54, 515-528.

[22] Dempster, A.P. (1968) "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society, Series B*, 30, 205-247.

[23] Denneberg, D. (1994a) *Non-additive Measures and Integral.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[24] Drèze, J.H. (1974) "Bayesian Theory of Identification in Simultaneous Equations Models." In S.E. Fienberg and A. Zellner eds. *Studies in Bayesian Econometrics and Statistics, In Honor of Leonard J. Savage*, pp. 175-191. North-Holland, Amsterdam.

[25] Epstein, L.G. and K. Seo (2012) "Bayesian Inference and Non-Bayesian Prediction and Choice: Foundations and an Application to Entry Games with Multiple Equilibria," mimeo, Boston Univesity and Northwestern University.

[26] Florens, J.-P. and M. Mouchart (1977) "Reduction of Bayesian Experiments." CORE Discussion Paper 7737, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium.

[27] Florens, J.-P., M. Mouchart, and J.-M. Rolin (1990) *Elements of Bayesian Statistics.* Marcel Dekker, New York.

[28] Florens, J.-P. and A. Simoni (2011) "Bayesian Identification and Partial Identification." mimeo, Toulouse School of Economics and Universita Bocconi.

[29] Galichon, A. and M. Henry (2006) "Inference in Incomplete Models," unpublished manuscript, Ecole Polytechnique and Universite de Montreal.

[30] Galichon, A. and M. Henry (2009) "A Test of Non-identifying Restrictions and Confidence Regions for Partially Identified Parameters" *Journal of Econometrics*, Vol. 152, 2, 186-196.

[31] Gustafson, P. (2009) "What Are the Limits of Posterior Distributions Arising from Nonidentified Models, and Why Should We Care?" *Journal of the American Statistical Association*, 104, No. 488, 1682-1695.

[32] Gustafson, P. (2010) "Bayesian Inference for Partially Identified Models," *The International Journal of Biostatistics*, 6, Issue 2.

[33] K. Hirano and J. Porter (2011) "Impossibility Results for Nondifferentiable Functionals," *forthcoming in Econometrica.*

[34] Horowitz, J., and C.F. Manski (2000) "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77-84.

[35] Huber, P.J. (1973) "The Use of Choquet Capacities in Statistics," *Bulletin of International Statistical Institute*, 45, 181-191.

[36] Hurwicz, L. (1950) "Generalization of the Concept of Identification," *Statistical Inference in Dynamic Economic Models, Cowles Commision Monograph 10.* John Wiley and Sons, New York.

[37] Imbens, G. W. and J. D. Angrist (1994) "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

[38] Imbens, G. W. and C.F. Manski (2004) "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845-1857.

[39] Kadane, J.B. (1974) "The Role of Identification in Bayesian Theory." In S.E. Fienberg and A. Zellner eds. *Studies in Bayesian Econometrics and Statistics, In Honor of Leonard J. Savage*, pp. 175-191. North-Holland, Amsterdam.

[40] Kass, R.E. and L. Wasserman (1996) "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 92, No. 435, 1343-1369.

[41] Kudō, H. (1967) "On Partial Prior Information and the Property of Parametric Sufficiency." In L. Le Cam and J. Neyman eds. *Proceedings the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 251-265.

[42] Koopmans and Riersol (1950) "The Identification of Structural Characteristics," *Annals of Mathematical Statistics* , 21, 165-181.

[43] Leamer, E.E. (1982) "Sets of Posterior Means with Bounded Variance Priors," *Econometrica*, 50, 725-736.

[44] Lehmann, E.L. and J. P. Romano (2005) *Testing Statistical Hypotheses, Second edition.* Springer Science + Business Media, New York.

[45] Liao, Y. and W. Jiang (2010) "Bayesian Analysis in Moment Inequality Models," *The Annals of Statistics*, 38, 1, 275-316.

[46] Manski, C.F. (1989) “Anatomy of the Selection Problem,” *The Journal of Human Resources*, 24, No. 3, 343-360.

[47] Manski, C.F. (1990) “Nonparametric Bounds on Treatment Effects,” *The American Economic Review*, 80, No. 2, 319-323.

[48] Manski, C.F. (2003) *Partial Identification of Probability Distributions.* Springer-Verlag, New York.

[49] Manski, C.F. (2007) *Identification for Prediction and Decision.* Harvard University Press, Cambridge, Massachusetts.

[50] Manski, C.F. (2009) “Diversified Treatment under Ambiguity,” *International Economic Review*, 50, 1013-1041.

[51] Molchanov, I. (2005) *Theory of Random Sets.* Springer-Verlag.

[52] Moon, H.R. and F. Schorfheide (2011) “Bayesian and Frequentist Inference in Partially Identified Models,” *forthcoming in Econometrica.*

[53] Neath, A.A. and F.J. Samaniego (1997) "On the Efficacy of Bayesian Inference for Nonidentifiable Models," *The American Statistician*, 51, 3, 225-232.

[54] Norets, A. and X. Tang (2012) "Semiparametric Inference in Dynamic Binary Choice Models," unpublished manuscript. Princeton University and University of Pennsylvania.

[55] Pearl, J. (1995) “On the Testability of Causal Models with Latent and Instrumental Variables,” *Uncertainty in Artificial Intelligence*, 11, 435-443.

[56] Picci, G. (1977) “Some Connections Between the Theory of Sufficient Statistics and the Identifiability Problem,” *SIAM Journal on Applied Mathematics*, 33, 383-398.

[57] Poirier, D.J. (1998) “Revising Beliefs in Nonidentified Models,” *Econometric Theory* , 14, 483-509.

[58] Rios Insua, D., F. Ruggeri, and B. Vidakovic (1995) "Some Results on Posterior Regret Γ-minimax estimation," *Statistics and Decisions*, 13, 315-331.

[59] Romano, J.P. and A.M. Shaikh (2010) "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169-211.

[60] Rothenberg, T.J. (1971) "Identification in Parametric Models," *Econometrica*, 39, No. 3, 577-591.

[61] Savage, L. (1951) "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46, 55-67.

[62] Shafer, G. (1976) *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

[63] Shafer, G. (1982) "Belief Functions and Parametric Models," *Journal of the Royal Statistical Society, Series B*, 44, No. 3, 322-352.

[64] Schennach, S.M. (2005) "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika*, 92, 1, 31-46.

[65] Schervish, M.J. (1995) *Theory of Statistics.* Springer-Verlag, New York.

[66] Song, K. (2012) "Point Decisions for Interval Identified Parameters," unpublished manuscript, University of British Columbia.

[67] Stoye, J. (2009) "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299-1315.

[68] van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge University Press.

[69] Vidakovic, B. (2000) "Γ-minimax: A Paradigm for Conservative Robust Bayesians." In D. Rios Insua and F. Ruggeri eds. *Robust Bayesian Analysis, Lecture Notes in Statistics*, 152, 241-259. Springer, New York.

[70] Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

[71] Wasserman, L.A. (1989) "A Robust Bayesian Interpretation of Likelihood Regions," *The Annals of Statistics*, 17, 3, 1387-1393.

[72] Wasserman, L.A. (1990) "Prior Envelopes Based on Belief Functions," *The Annals of Statistics*, 18, 1, 454-464.

[73] Wasserman, L.A. and J.B. Kadane (1990) "Bayes Theorem for Choquet Capacities," *The Annals of Statistics*, 18, 3, 1328-1339.